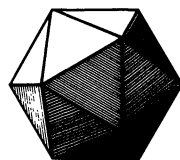


The American Mathematical Monthly



Volume 102, Number 1 / JANUARY 1995



Dividing a Cake
(see page 9)

NOTICE TO AUTHORS

The *Monthly* publishes articles, notes, and other features about mathematics and the profession. The readership of the *Monthly* is intended to include everybody who is mathematically inclined, including of course professional mathematicians and students of mathematics at all collegiate levels. While no single article or feature is likely to appeal to everyone, material should interest and be accessible to a large number of readers. This is the most important criterion for acceptance.

Articles may be expositions of old results or presentations of new ones. They may concern all of mathematics or one small area, a broad development or a single application, historical reminiscences or one important event. While some articles may contain the author's new research, the novelty of material and generality of the results is far less important than the clarity of exposition and general interest. Discussing one illuminating case of a well known result is far better than providing all the details of an obscure but new proposition. Articles in the *Monthly* are supposed to inform and to entertain; they are meant to be read rather than archived.

Notes are short and possibly informal articles. A note may concern a clever new proof of an old theorem, a novel way to present tired material, or a lively discussion of a philosophical (but still mathematical) issue. Also, any topic is suitable, so long as it is related to mathematics. Because a note is short, the first few sentences are the most important part: They should explain the purpose and invite the reader in. Photographs or diagrams often will attract the reader's attention.

All articles and notes should be sent to the editor:

JOHN EWING
Department of Mathematics
Indiana University
Bloomington, IN 47405

Please send 3 copies, typewritten on only one side of the paper. Illustrations should be carefully drawn on separate sheets of paper in black ink; the original should be without lettering and two copies should have appropriate captions and lettering indicated.

Proposed problems or solutions should be sent to:

RICHARD BUMBY,
P.O. Box 10971
New Brunswick, NJ 08906-0971.

Please send 2 copies of all material, typewritten if possible.

Letters to the Editor, both for publication and for private reading, should be sent to the Editor at the address given above. Comments, including criticisms, are welcome, as are all suggestions for making the *Monthly* a lively, entertaining, and informative journal.

EDITOR:

JOHN H. EWING

ASSOCIATE EDITORS:

PETER BORWEIN
RICHARD BUMBY
DENNIS DETURCK
UNDERWOOD DUDLEY
JOHN DUNCAN
JOAN FERRINI-MUNDY
JOSEPH GALLIAN
STEVEN GALOVICH
RICHARD GUY
DARRELL HAILE
PAUL HALMOS
JOAN HUTCHINSON

FRED KOCHMAN
CATHERINE MCGEOCH
RICHARD NOWAKOWSKI
ARNOLD OSTEBEE
LEE RUBEL
ABE SHENITZER
LYNN STEEN
STAN WAGON
DOUGLAS WEST
HERBERT WILF
SANDY ZABELL
PAUL ZORN

EDITORIAL ASSISTANT:

MISTY CUMMINGS

STAFF ARTIST:

MIKE CAGLE

QUOTE MASTER:

MARK WOODARD

Reprint permission:

MARCIA P. SWARD, Executive Director

Advertising Correspondence:

Ms. ELAINE PEDREIRA, Advertising Manager

Subscription correspondence, change of address,
and other inquiries:

Membership / Subscriptions Department

All at the address:

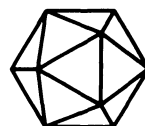
The Mathematical Association of America
1529 Eighteenth Street, N.W.
Washington, DC 20036.

Microfilm Editions: University Microfilms International,
Serials coordinator, 300 North Zeeb Road, Ann
Arbor, MI 48106.

The AMERICAN MATHEMATICAL MONTHLY (ISSN 0002-9890) is published monthly except bimonthly June-July and August-September by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, DC 20036 and Montpelier, VT. Copyrighted by the Mathematical Association of America (Incorporated), 1995, including rights to this journal issue as a whole and, except where otherwise noted, rights to each individual contribution. General permission is granted to Institutional Members of the MAA for noncommercial reproduction in limited quantities of individual articles (in whole or in part) provided a complete reference is made to the source. Second class postage paid at Washington, DC, and additional mailing offices. **Postmaster:** Send address changes to the American Mathematical Monthly, Membership / Subscription Department, MAA, 1529 Eighteenth Street, N.W., Washington, DC, 20036-1385.

The American Mathematical Monthly

Volume Volume 102, Number 1 / JANUARY 1995
(ISSN 0002-9890)



Contents

ARTICLES

What Is the Worth of Free Casino Credit? / MICHAEL ORKIN
and RICHARD KAKIGI 3

An Envy-Free Cake Division Protocol / STEVEN J. BRAMS
and ALAN D. TAYLOR 9

The Mathematics Portfolio / MARY L. CROWLEY
and KEN DUNN 19

Derivative Polynomials for Tangent and Secant /
MICHAEL E. HOFFMAN 23

The Law of Large Numbers and $\sqrt{2}$ / THOMAS M. LIGGETT
and PETER PETERSEN 31

Exponentiation in Rings / R. H. REDFIELD 36

Integer Hexahedra Equivalent to Perfect Boxes / BLAKE E. PETERSON
and JAMES H. JORDAN 41

FEATURES

COMMENTS 2

NOTES

Calculating Normal Probabilities / RICHARD J. BAGBY 46

Constrained Critical Points / PAUL SHUTLER 49

A Cone Eversion / S. TABACHNIKOV 52

THE COMPUTER SCIENCE SAMPLER

Approximation Algorithms: Good Solutions to Hard Problems /
RAN LIBESKIND-HADAS 57

THE EVOLUTION OF...

Four Significant Axiomatic Systems and Some of the Issues Associated
with Them / STEFAN MYKYTIUK and ABE SHENITZER 62

THE AUTHORS 68

PROBLEMS AND SOLUTIONS 70

REVIEWS

Algebra, by I. M. Gelfand and Alexander Shen / RICHARD ASKEY 78

TELEGRAPHIC REVIEWS 82

What is the Worth of Free Casino Credit?

Michael Orkin and Richard Kakigi

THE ZARIN CASE—In 1980, a compulsive gambler named David Zarin used a generous credit line to run up a huge debt playing craps in an Atlantic City casino. When the casino finally cut off Zarin's credit, he owed over \$3 million. Due in part to New Jersey's laws protecting compulsive gamblers, the debt was deemed unenforceable by the courts, leading the casino to settle with Zarin for a small fraction of the amount he owed. Later, the Internal Revenue Service tried to collect taxes on the approximately \$3 million Zarin didn't repay, claiming that cancellation of the debt made it taxable income. Since Zarin had never actually received any cash (he was always given chips, which he promptly lost at the craps table), an appellate court finally ruled that Zarin had no tax obligation [6]. The courts never asked what Zarin's credit line was actually worth. Surely, it was worth something. With \$3 million dollars in chips to play with, there is a chance, albeit small, that a gambler will make a profit and leave the casino with cash in his pocket. We will show that, viewed as a gambler's ruin problem making pass line bets at craps, the "worth" of a sufficiently large free line of credit is approximately \$197,000.

FREE CREDIT—Suppose a casino gives you \$3 million in chips for gambling, but to redeem chips for cash you must first pay back the \$3 million. You incur no debt if you lose. You restrict your betting to the pass line bet in craps. What is your optimal betting strategy and how much profit will it yield on the average?

To answer this question, we must define "optimal." According to the laws of probability, in repeated play of a game in which the house has an edge, you will eventually end up broke no matter *what* strategy you use. We have to settle for something more modest than guaranteed success.

BOLD PLAY—Consider the following approach. You have a fixed monetary goal and will keep betting until you either reach your goal, in which case you quit a winner, or go broke. Nothing else matters. For you, a strategy is optimal if it maximizes the chance that you reach your goal. Dubins and Savage showed that when the house has an edge, *bold play* is optimal. Bold play means that you bet your entire bankroll on each bet, or as much as is necessary to reach your goal. For example, suppose you have a bankroll of \$9,000 and your goal is \$20,000. On your first bet, bet the entire \$9,000. If you lose, you are broke, and the game is over. If you win, your bankroll has increased to \$18,000. Since you are now only \$2,000 short of your goal, bet \$2,000. If you win, you have reached your goal. If you lose, your bankroll is \$16,000. Bet \$4,000, just enough to reach your goal. And so on. Caution: Bold play doesn't guarantee a profit or shift the edge to your favor. It only maximizes the probability that you will reach your goal before going broke.

If you are a bold player *and* a big bettor, you will have a problem if the necessary bet exceeds the house limit. For example, suppose you have \$10,000, your goal is \$25,000, and the maximum bet is \$5,000. Bold play says bet \$10,000, but because of the betting limit, you are forced to bet \$5,000. When you always bet the same amount, the chance of reaching your goal can be computed with the Gambler's Ruin formula.

GAMBLER'S RUIN—Suppose you bet repeatedly, always betting the same amount, which we shall call a “betting unit,” and suppose you keep playing until you either reach your goal or go broke. Then, assuming 1 to 1 payoff odds, the chance that you reach your goal is computed as follows.

Let

p = your chance of winning one bet

$q = 1 - p$ = your chance of losing one bet

i = your current fortune, expressed in betting units

N = your goal, expressed in betting units ($N > i$).

Then if $P(i, N) = \text{Prob}(\text{you reach } N \text{ before going broke})$ we have (Feller, page 344)

$$P(i, N) = \frac{1 - \left(\frac{q}{p}\right)^i}{1 - \left(\frac{q}{p}\right)^N} \quad \text{if } p \neq \frac{1}{2};$$

$$= \frac{i}{N} \quad \text{if } p = \frac{1}{2}.$$

GAMBLER'S RUIN AND THE PASS LINE BET—The pass line bet in craps is almost a fair game. The approximate chance of winning is .493 (the exact chance is 244/495). The approximate chance of losing is .507. Payoff odds are 1 to 1. According to the law of averages, in repeated play, a gambler will win 49.3% and lose 50.7% of his bets for a net loss of 1.4 cents per dollar bet. Suppose that you currently have \$10,000, your goal is \$25,000, and you always bet \$5,000 on the pass line. Then using the approximate probabilities

$$p = .493$$

$$q = .507$$

$$i = 2 \text{ (2 betting units of \$5,000)}$$

$$N = 5 \text{ (5 betting units)}$$

and

$$P(\text{you reach goal before going broke}) = \frac{1 - \left(\frac{.507}{.493}\right)^2}{1 - \left(\frac{.507}{.493}\right)^5} = .383.$$

The chance that you reach the goal of \$50,000 (10 betting units) before going broke is .18. The chance that you reach the goal of \$100,000, is .08. The chance of turning \$10,000 into \$1 million by repeatedly making \$5,000 pass line bets is .0002.

The bigger your goal, the less your chance of reaching it. In fact, for any bet where the chance of losing is greater than the chance of winning ($q/p > 1$), as the goal, N , gets larger, the chance of reaching it gets smaller, eventually approaching 0. When the house has an edge, repeated play grinds down the gambler's bankroll.

EXPECTED PROFIT—What are your average winnings with free casino credit when you always make maximum pass line bets of \$15,000? Say you start with i units credit and keep betting until you either reach N units or go broke ($N > i$). You will end up with either a profit of $N - i$ or 0 (no debt). Denoting Expected Profit by $E(i, n)$, we get

$$E(i, N) = (N - i)P(i, N) = (N - i) \frac{1 - \left(\frac{q}{p}\right)^i}{1 - \left(\frac{q}{p}\right)^N} \quad \text{if } p \neq \frac{1}{2}.$$

This amount varies according to the goal, N . Naturally, we want to find the best N , that is, the N that maximizes $E(i, N)$. Starting with \$3 million credit, always making maximum \$15,000 pass line bets, we have $i = 200$ units, $p = .493$, $q = .507$ and

$$E(200, N) = (N - 200) \times P(200, N) = (N - 200) \times \frac{1 - \left(\frac{.507}{.493}\right)^{200}}{1 - \left(\frac{.507}{.493}\right)^N}.$$

This expression reaches a maximum when $N = 236$ (36 unit profit), in which case, Expected Profit = \$197,000 (approximately). The optimal strategy, then, is to keep making maximum bets until you either reach a 36 unit (\$540,000) profit and quit, or go broke. Using this strategy, a \$3 million free credit is “worth” \$197,000. (It is not clear how the judge would have used this information, since tax liability is based upon actual rather than expected gains and losses.)

THERE IS A LIMIT—It turns out that there is a limit to what free credit can do. After a while, increasing the credit line doesn't add to Expected Profit. In fact, the Expected Profit of \$197,000 is the most a craps-playing gambler can hope for, no matter what the credit line. If you happen to get \$1 billion free credit, or even \$100 billion, you should still play for a 36 unit profit, and your Expected Profit is still \$197,000. This interesting fact is obtained as follows.

Let $N \rightarrow \infty$ and $i \rightarrow \infty$ with $D = N - i$ fixed, and let $r = q/p > 1$. Then

$$P(i, N) = \frac{1 - r^i}{1 - r^N} \rightarrow r^{-D}$$

and $E(i, N) \rightarrow Dr^{-D}$. The maximum of Dr^{-D} occurs when

$$D = \frac{1}{\ln(r)}.$$

Thus, for large i , the target increment that maximizes expected payoff is approximately $1/\ln(r)$. Using this increment for each i , and letting $i \rightarrow \infty$, we have

$$P\left(i, i + \frac{1}{\ln(r)}\right) \rightarrow r^{-1/\ln(r)} = e^{-1} \quad \text{and} \quad E\left(i, i + \frac{1}{\ln(r)}\right) \rightarrow (e \ln(r))^{-1},$$

where e is the base for natural logarithms ($e = \text{approx } .37$). If one unit = \$C, then for large i , the worth of i units of credit is approximately $\$C/(e \ln(r))$.

For pass line bets (using the approximate probability $p = .493$), $r = .507/.493 = 1.028$, so $\ln(r) = .028$ and $D = 1/.028 = 36$. Thus, for large i , the optimal goal is a 36 unit profit and your chance of reaching it before going broke is $1/e = .37$ (In fact, for large i , your chance of reaching the optimal goal before going broke is $1/e$, REGARDLESS of the game you are playing, as long as $q > p$). In this case, Expected Profit = $1/e(0.028) = 13.14$ units. When the betting unit = \$15,000, Expected Profit = \$197,000.

FAVORABLE GAMES—Bold play is fine if the house has an edge, but what if YOU have an edge? This doesn't happen in most casino games, but it is generally agreed that blackjack, sports betting, and horse racing can, at least theoretically, give a clever player an edge. In this case, the law of averages guarantees that the gambler, not the casino, will be a winner in repeated play, provided that reckless betting doesn't cause early bankruptcy. When the gambler has a huge bankroll, even maximum allowable bets are prudent, and there is little chance of going broke. In this case, the gambler is almost assured of whatever goal he wants to reach, subject to various constraints. Consider blackjack.

BLACKJACK—Skilled blackjack players will win about 51% of their bets under ideal playing conditions (It is unclear how often such conditions exist, especially for big bettors). Suppose the maximum bet is \$15,000, same as craps, and the gambler, having a \$3 million bankroll, makes \$15,000 blackjack bets with win probability = .51. Applying the gambler's ruin formula, we see a stark difference between blackjack and craps.

In craps (or any bets where your chance of losing is greater than your chance of winning), your chance of reaching your goal before going broke decreases to 0 as your goal increases. In blackjack, with $p = .51$, (or any bets where your chance of winning is greater than your chance of losing), your chance of reaching your goal before going broke decreases as your goal increases, but *not* to 0. Instead, starting with i betting units, your chance of reaching your goal never goes below $1 - (q/p)^i$. With $p = .51$, $q = .49$, and starting with $i = 200$ betting units, your chance of reaching *any* goal before going broke is always greater than .999. This follows directly from the gambler's ruin formula. With a 200 betting unit bankroll, betting a unit at a time, you're almost certain to reach your goal, no matter how large, provided that you have enough time to play (if you set your sights unrealistically high, you may die of old age before you reach your goal).

WHAT A DIFFERENCE AN EDGE MAKES—As an example of how the slight difference in pass line and blackjack win probabilities is magnified over repeated play, suppose you start with $i = 200$ betting units and have a goal of $N = 500$ units. Making pass line bets, with $p = .493$, your chance of reaching 500 units before going broke is .00023. Making blackjack bets, with $p = .51$, your chance of reaching 500 units is .99966.

RECKLESS BANKRUPTCY—If your chance of winning is greater than your chance of losing, you can almost certainly reach your goal with prudent play. However, if your betting unit is recklessly large relative to your bankroll, you may suddenly go broke. For example, suppose you are playing blackjack and you start by betting your entire bankroll. With win probability = .51, you have a .49 chance of going broke on the first bet. If there is no betting limit, and you always bet your entire bankroll, you are almost certain to go broke at some point. Thus, you shouldn't bet recklessly.

THE KELLY SYSTEM —The Kelly system, named after the mathematician J. L. Kelly who invented it, specifies how much to bet if you have an edge, as in the case of an expert blackjack player. With the Kelly system, you always bet a fixed fraction of your bankroll. The more you have, the more you bet, the less you have, the less you bet. Leo Breiman showed that the Kelly system is optimal for favorable games (you have an edge) in two important ways: First, it does better in the long run than any substantially different strategy. Second, the expected number of bets necessary to reach a specified goal with the Kelly system is less than with any other strategy. If the house has an edge, as with the pass line bet, the Kelly system says “don't bet.”

To use the Kelly system, you must know your chance of winning and the payoff odds. Mathematically speaking, you bet the fraction of your bankroll that maximizes “growth rate.” Growth rate is the expected logarithm of your return, where return is payoff per dollar bet.

If the payoff odds are 1 to 1 and p is your chance of winning, you should always bet the fraction $2 \times p - 1$ of your bankroll. For example, with blackjack win probability = .51, the Kelly system says bet the fraction $2 \times .51 - 1 = .02$ of your bankroll. If your bankroll is \$500, bet \$10. If your bankroll is \$5,000, bet \$100. If you have \$3 million, bet \$60,000. If you have \$3 million and the betting limit is \$15,000, bet the limit. (Note: This is a simplification. Good blackjack play requires card counting, with bet increases when the deck favors the player).

HOURLY WAGE—For any betting strategy, it may be useful to compute the ratio of the expected profit to the expected duration of play. This will give an estimate of your expected hourly wage. To do this, you need to know the average number of bets in a betting sequence. In the gambler's ruin sequence, the average number of bets it takes to either reach your goal or go broke is given by the formula (Feller, page 348)

$$B = \frac{i}{(q - p)} - \left(\frac{N}{(q - p)} \right) \left(\frac{1 - (q/p)^i}{1 - (q/p)^N} \right).$$

For example, if you are playing blackjack, with $p = .51$ and $q = .49$, starting with the usual $i = 200$ betting units, and having the goal of $N = 500$ units, then $B = 14,992$. If you make an average of 1 bet per minute, it will take about 250 hours, or about one month of 8 hour day, 6 day weeks to either reach your goal or go broke. Suppose, like Zarin, you have \$3 million, and, unlike Zarin, you make \$15,000 blackjack bets. Then, your Expected Profit is about \$4.5 million, whether or not you have free casino credit. Dividing by 250 hours yields an hourly wage of \$18,000. Not bad. Unfortunately, most of us don't have a \$3 million gambling bankroll (or the ability to win a casino game 51% of the time). If you start with

\$3,000 instead of \$3 million and make \$15 blackjack bets instead of \$15,000 bets, your average hourly wage will be \$18 instead of \$18,000.

Let's go back to Zarin and his \$3 million, expressed as 200 pass line betting units of \$15,000. It will take an average of 8149 bets, or about 136 hours, at one bet per minute, to either reach the 236 unit goal or go broke. We saw that with free credit, Zarin's Expected Profit from using this strategy was \$197,000. His expected hourly wage would have been $\$197,000/136 = \1449 .

WHAT DOES BANKROLL REALLY MEAN?—Say you bring \$1,000 to the casino. Is that your bankroll? Suppose you have another \$10,000 in the bank. Should that be counted? You can sell your car. Is that part of your bankroll? Maybe you can also sell your house and milk your credit cards to the limit. There are many possibilities, but if you bloat your bankroll with money not meant for gambling, losing can be painful.

BOTTOM LINE—Zarin should have learned blackjack. If he had mastered a strategy giving him a slight edge (such as win probability = .51) and if he had been allowed to make large bets under ideal playing conditions, his \$3 million credit would have allowed him to generate considerable profit. Unfortunately, ideal playing conditions are hard to come by for big-betting blackjack players.

ACKNOWLEDGMENT—The authors would like to thank Richard Beck of the New York Law School for informing them of the Zarin case.

REFERENCES

1. L. Breiman (1961), Optimal Gambling Systems for Favorable Games, *Fourth Berkeley Symposium on Probability and Statistics*, I, 65–78.
2. L. Dubins, L. J. Savage (1965), *How to Gamble If You Must*, McGraw-Hill, New York.
3. W. Feller (1968), *An Introduction to Probability Theory and Its Applications*, Vol. I, 3rd Edition, John Wiley, New York.
4. J. L. Kelly (1956), A New Interpretation of Information Rate, *Bell System Technical Journal*, 35, 917–926.
5. M. Orkin (1991), *Can You Win?*, W. H. Freeman and Co, New York.
6. David and Louise Zarin vs. Commissioner of Internal Revenue; David and Louise Zarin, Appellants, No. 90-1240, US Court of Appeals, for the Third Circuit.

Department of Statistics
California State University, Hayward
Hayward, CA 94542-3087
morkin@csu Hayward.edu
rkakigi@csu Hayward.edu

There is no national science just as there is no national multiplication table; what is national is no longer science.

—Anton Chekov (1860–1904)

W. H. Auden and L. Kronenberger, eds. *The Viking Book of Aphorisms*, New York: The Viking Press, 1966.

An Envy-Free Cake Division Protocol

Steven J. Brams and Alan D. Taylor

Our starting point is the well-known parental solution to the problem of dividing a cake between two children so that each child thinks he or she has been treated fairly: The parent instructs one of the children to cut the cake into two pieces in any way he desires. The other child is then instructed to choose whichever piece she prefers. This two-step sequence of instructions, known as “cut-and-choose,” provides a simple example of the kind of game-theoretic algorithm that Even and Paz [11] call a “protocol.”

Associated with the cut-and-choose protocol is a natural strategy for each child: The first child cuts the cake into two pieces that he considers to be equal, and the second child chooses a piece that she considers to be at least as large as the other piece. Notice that each child’s strategy guarantees him or her “satisfaction,” regardless of what the other child does.

The general version of this problem involves n people (“players”), each of whom has his or her preferences over subsets of the cake given by a probability measure.¹ An allocation of the cake among the players is said to be *proportional* if each player receives a piece of size at least $1/n$ (in his or her own measure), and it is said to be *envy-free* if each player receives a piece he or she would not swap for that received by any other player. It is easy to see that an envy-free allocation is proportional, but the converse fails unless $n = 2$. Thus, for example, every one of three players may think his or her piece is at least $1/3$, but a player may think that one of the other players has a larger piece.

The results on proportional and envy-free allocations obtained over the past 50 years tend to fall into one of four classes: (i) Existence Theorems; (ii) Moving-Knife Solutions; (iii) Algorithms; and (iv) Protocols. We say something about each in turn.

Existence theorems, dating back to the 1940s, are often based on some version of Liapounoff’s Convexity Theorem [20]. Typically, they establish the existence of an ordered partition of the cake corresponding to an envy-free allocation, often with some additional property such as: all the measures of all the pieces are exactly $1/n$ [21]; or the pieces are connected sets [27 and 31]; or the allocation is also Pareto-optimal [30 and 4]. In the words of Rebman [24, p. 33], however, these results provide “no clue as to how to accomplish such a wonderful partition.”

There are two well-known moving-knife solutions. The first is due to Dubins and Spanier [10] and is a moving-knife version of the Banach-Knaster last-

¹If one wants to abandon the cake metaphor, and literally work with arbitrary probability measures on some set C , then even cut-and-choose can fail. For example, if both children have their preferences given by the same 0-1 valued measure, i.e., the same ultrafilter, then both will want the same piece regardless of how the cake is divided.

For everything we do in the present paper, it suffices to assume that our measures are all defined on the same algebra \mathcal{A} of subsets of C and satisfy the following two properties: (i) for every set $P \in \mathcal{A}$ and every finite k , P can be partitioned into k sets of equal measure, and (ii) for every $P, Q \in \mathcal{A}$, either P can be trimmed to yield a subset the same size as Q , or vice-versa.

diminisher scheme for n people that we shall present later. (The Dubins-Spanier scheme is easy to describe: a knife is slowly moved along the top of the cake so that all the slices made are parallel. Each player calls “cut” when he or she is willing to take the resulting piece as his or her allocation.) The second is a moving-knife scheme due to Stromquist [27] that yields an envy-free allocation among three people. (This one is not so easy to describe, because envy-freeness is considerably harder to obtain than proportionality.)

Moving-knife schemes, however, are not the step-by-step processes one usually associates with the term “algorithm.” A good example of what one would call an algorithm in this context is Woodall’s scheme [32] for producing an allocation whereby each participant gets strictly more than $1/n$ of the cake (according to his own measure). This algorithm requires, as part of the input, a piece P of cake and two distinct numbers α and β such that Player 1 thinks the measure of P is α , and Player 2 thinks the measure of P is β . An envy-free version of this algorithm is in [5].

Finally, there is what Even and Paz [11] call a “protocol,” and this is the only kind of result we are going to analyze in the present paper. Since we will be presenting examples of protocols, as opposed to proving their non-existence, we can afford the same level of informality in the description of what is meant by a protocol as Even and Paz used.

A *protocol* is a computer-programmable interactive procedure that can issue queries to the participants whose answers may affect future decisions. It may issue instructions to the participants such as: “Choose k pieces from among these m pieces” or “Partition this piece into k subpieces.” The protocol has no information on the measure of the various pieces as seen by the different participants—this is private information. Moreover, if the participants obey the protocol, then each participant will end up with a piece after finitely many steps.

Still following [11], we define a *strategy* for a participant to be an adaptive sequence of moves consistent with the protocol, which the participants choose sequentially when called upon by the protocol. A protocol is *proportional* if each of the n participants has a strategy that will guarantee him at least $1/n$ of the cake (by his own measure), independently of the other participants’ strategies. (Purely for convenience, we will henceforth use only the masculine pronoun.) Departing from [11], we will call a protocol *envy-free* if each of the n participants has a strategy that will guarantee him a piece that is, according to his own measure, at least tied for largest.

A constructive proof of the existence of, say, a proportional protocol involves producing three separate things: the rules of the protocol, a strategy for each of the players, and an argument that the strategies do, in fact, guarantee each player his proportional share. We distinguish rules and strategies by demanding that rules be enforceable by a referee implementing the protocol.

This means that a statement like “Player 1 cuts the cake into n pieces” is an acceptable rule, whereas a statement like “Player 1 cuts the cake into n pieces that he considers to be equal” is not. This is because the latter statement cannot be enforced by the referee, who has no knowledge of Player 1’s measure and so cannot tell if the rule has been followed or not.

In presenting protocols, we will separate rules from strategies by placing all strategic aspects in parentheses. This provides one with the option of reading the rules alone in a reasonably smooth way. All arguments that the strategies perform as advertised are placed between steps and labeled as “Aside.” For example, in our method of presentation, cut-and-choose becomes:

Cut-and-Choose

- Step 1. Player 1 cuts the cake into 2 pieces (that he considers to be the same size).
- Step 2. Player 2 chooses a piece (that she considers to be at least tied for largest).
- Aside. Clearly, Player 1's strategy guarantees him a piece of size exactly $1/2$ in his measure, while Player 2's strategy guarantees her a piece of size at least $1/2$ in her measure.

The modern era of cake cutting began with Steinhaus' observation "during the war [World War II]" [25, p. 102] that the cut-and-choose protocol could be extended to yield a proportional protocol for three players (see [18]). He then asked if it could be extended to yield a proportional protocol for the case $n > 3$. (Steinhaus, however, never used the word "protocol.") His question was answered in the affirmative by Banach and Knaster and reported in [25] and [26]. The Steinhaus and Banach–Knaster protocols introduced two key ideas that would resurface in the envy-free solutions 15 and 50 years later.

The first idea was that of having an initial sequence of steps resulting in only part of the cake's being allocated (to one player in this case). The sequence is then repeated a finite number of times, after which the entire cake has been allocated. The second idea—and perhaps the more important of the two—was that of having a player trim a piece to a smaller size.

Explicit mention of the lack of a *constructive* procedure for producing an envy-free allocation among more than two people dates back at least to Gamow and Stern [14]. The first breakthroughs on this problem occurred in the late 1950s and early 1960s, when the protocol solution to the envy-free problem for $n = 3$ was found by John L. Selfridge, and rediscovered independently by John H. Conway. These solutions also involved trimming and an initial allocation of only part of the cake; they were widely disseminated by R. K. Guy and others, and eventually reported by Gardner [15], Woodall [32], Stromquist [27], and Austin [1]. The moving-knife solution of Stromquist [27] was found two decades later, as was a scheme due to Levmore and Cook [19], which can be recast as quite a different moving-knife solution to the envy-free problem when $n = 3$. Still other envy-free moving-knife schemes for three people [7] and, more recently, four people [8] have been discovered and are summarized in [6].

The extension of the Selfridge–Conway protocol to the case of even four people has remained an open, and much-commented upon, problem. See, for example, Gardner [15], Rebman [24], Stromquist [27], Woodall [31], [32], Bennett *et al* [3], Webb [29], Hill [16], [17], and Olivastro [22]. In what follows, we solve this problem by producing an envy-free protocol for arbitrary n .

We have chosen a uniform presentation of four protocols that highlights the evolution of two important ideas—namely, trimming, and the use of sequences of partial allocations. Historically, these four protocols arose over a period of 50 years and nicely illustrate how ideas in mathematics are built, one upon another. The protocols we present are:

1. The proportional protocol for $n = 3$ (Steinhaus).
2. The proportional protocol for arbitrary n (Banach–Knaster).
3. The envy-free protocol for $n = 3$ (Selfridge, Conway).
4. The envy-free protocol for arbitrary n .

Before turning to the protocols themselves, we must acknowledge the help of several people. Our interest in fair division was sparked by Olivastro [22]. Valuable mathematical contributions were made by William Zwicker and Fred Galvin. Indeed, the present version of our envy-free protocol owes much to the reworking of an earlier version by Galvin.

Specific observations and comments by David Gale, Sergiu Hart, Theodore Hill, Walter Stromquist, William Webb, and Douglas Woodall also proved helpful. In addition, we have benefited from conversations and correspondence with the following people: Ethan Akin, Julius Barbanel, John Conway, Morton Davis, Karl Dunz, Shimon Even, A. M. Fink, Peter Fishburn, Martin Gardner, Richard Guy, D. Marc Kilgour, Peter Landweber, Jerzy Legut, Hervé Moulin, Dominic Olivastro, Barry O'Neill, Philip Reynolds, William Thomson, Hal Varian, Charles Wilson, and H. Peyton Young.

The first protocol we present is a generalization of cut-and-choose to a proportional protocol for three people. This is the one found by Steinhaus during World War II.

The Proportional Protocol for $n = 3$

(Steinhaus, circa 1943)

- Step 1. Player 1 cuts the cake into 3 pieces (that he considers to be the same size).
- Step 2. Player 2 is given the choice of either passing, i.e., doing nothing (which he does if he thinks 2 or more of the pieces are of size at least $1/3$), or not passing and labeling 2 of the pieces (that he thinks are of size strictly less than $1/3$) as “bad.”
- Step 3. If Player 2 passed in step 2, then Players 3, 2, and 1, in that order, choose a piece (that they consider to be of size at least $1/3$).
- Aside. In this case, each player receives a piece of size at least $1/3$ in his own measure. This is true of: Player 3, because he chooses first; Player 2, because he thinks either 2 or 3 pieces are that large, and so at least one of them will still be available after Player 3 chooses his piece; and Player 1, because he made all 3 pieces of size $1/3$.
- Step 4. If Player 2 did not pass at Step 2, then Player 3 is given the same two options that Player 2 had at Step 2. He ignores Player 2's labels.
- Step 5. If Player 3 passed in Step 4, then Players 2, 3, and 1, in that order, choose a piece (that they consider to be of size at least $1/3$).
- Aside. In this case, as before, each player receives a piece of size at least $1/3$ in his own measure.
- Step 6. If Player 3 did not pass at Step 4, then Player 1 is required to take a piece that both Player 2 and Player 3 labelled as “bad.”
- Aside. Note first that there certainly must be such a piece. At this point, Player 1 has received a piece that he thinks is of size exactly $1/3$, which both Player 1 and Player 2 think is “bad,” i.e., of size strictly less than $1/3$.
- Step 7. The other two pieces are reassembled, and Player 2 cuts the resulting piece into two pieces (that he considers to be the same size).
- Step 8. Player 3 chooses one of the two pieces (that he considers to be at least tied for largest).
- Step 9. Player 2 is given the remaining piece.
- Aside. This is just cut-and-choose between Players 2 and 3, which ends the protocol.

The second protocol we present followed quickly on the heels of the first. It is the Banach-Knaster protocol, offered in response to Steinhaus' question of whether his result could be extended from 3 to n people. Note here the introduction of the idea of trimming, which will be further exploited in both of the upcoming envy-free protocols.

Proportional Protocol for Arbitrary n
(Banach-Knaster, circa 1944)

- Step 1. Player 1 cuts a piece P_1 (of size $1/n$) from the cake.
 - Step 2. Player 2 is given the choice of either passing (which he does if he thinks P_1 is of size less than $1/n$), or trimming a piece from P_1 to create a smaller piece (that he thinks is of size exactly $1/n$). The piece P_1 , now perhaps trimmed, is renamed P_2 . The trimmings are set aside.
 - Step 3. For $3 \leq i \leq n$, Player i takes the piece P_{i-1} and proceeds exactly as Player 2 did in Step 2, with the resulting piece now called P_i .
 - Aside. For $1 \leq i \leq n$, Player i thinks that P_i is of size less than or equal to $1/n$. We also have that $P_1 \supset \cdots \supset P_n$. Thus, every player thinks P_n is of size at most $1/n$.
 - Step 4. The last player to trim the piece, or Player 1 if no one trimmed it, is given P_n .
 - Aside. The player receiving P_n thinks it is of size exactly $1/n$.
 - Step 5. The trimmings are reassembled, and Steps 1–4 are repeated for the remainder of the cake, and with the remaining $n - 1$ players in place of the original n players.
 - Aside. The player who gets a piece at this second stage is getting exactly $1/(n - 1)$ of the remainder of the cake; he, and everyone else, thinks this remainder is of size at least $(n - 1)/n$. Hence, he thinks his piece is of size at least $1/n$.
 - Step 6. Step 5 is iterated until there are only 2 players left. The last 2 players use cut-and-choose.
 - Aside. As before, each player receives a piece that he thinks is of size at least $1/n$.
- This ends the protocol.

The next protocol we present is the Selfridge-Conway envy-free protocol for the case $n = 3$. (There are slight differences in the presentations of Selfridge and Conway; we follow the latter.) This protocol involves an elegant combination of the trimming idea introduced by Banach-Knaster and the basic framework that Steinhaus used. It also introduces the important notion of one player's having an "irrevocable advantage" over another player, following a partial allocation.

Envy-Free Protocol for $n = 3$
(Selfridge, Conway, circa 1960)

- Step 1. Player 1 cuts the cake into 3 pieces (that he considers to be the same size).
- Step 2. Player 2 is given the choice of either passing (which he does if he thinks two or more pieces are tied for largest), or trimming a piece from (the largest) one of the three pieces (to create a tie for largest). If Player 2 trimmed a piece, then the trimmings are named L , for "leftover," and set aside.

- Step 3. Players 3, 2, and 1, in that order, choose a piece (that they consider to be at least tied for largest) from among the 3 pieces, one of which may have been trimmed in Step 2. If Player 2 did not pass in Step 2, then he is required to choose the piece he trimmed if Player 3 did not.
- Aside. Notice that only part of the cake has been allocated. This yields a partition $\{X_1, X_2, X_3, L\}$ of the cake such that $\{X_1, X_2, X_3\}$ is an envy-free partial allocation. The lack of envy is true of: Player 3, because he chooses first; Player 2, because he made at least two pieces tied for largest, and so at least one of them will still be available after Player 3 chooses his piece; and Player 1, because he made all three pieces of size $1/3$, and the trimmed one has definitely been taken by either Player 3 or Player 2.
- Step 4. If Player 2 passed at Step 2, we are done. Otherwise, either Player 2 or Player 3 received the trimmed piece, and the other received an untrimmed piece. Whichever player received the *untrimmed* piece now divides L into 3 pieces (that he considers to be the same size). Call this player the “cutter” and the other the “non-cutter.”
- Aside. We will refer to Player 1 as having an *irrevocable advantage* over the non-cutter. The point is that, since the non-cutter received the trimmed piece, Player 1 will not envy the non-cutter, *regardless* of how L is later divided among the three.
- Step 5. The three pieces into which L is divided are now chosen by the players in the order: non-cutter first; Player 1 second; cutter third. (Each chooses a piece at least tied for largest among those available to him when it is his turn to choose.)
- Aside. At this point, the entire cake has been allocated. Since the non-cutter chooses his piece of L first, he experiences no envy. Player 1 does not envy the non-cutter, since he had an irrevocable advantage over him, and Player 1 does not envy the cutter, because he is choosing his piece of L before the cutter does. Finally, the cutter experiences no envy since he divided L into three equal pieces.

This ends the protocol.

The final protocol we present is our envy-free protocol for an arbitrary number of players. This result was announced in [9, 12, 13, 23]. A brief discussion of some important differences between this protocol and the three earlier ones, and a couple of important open questions, follow.

The central feature of our envy-free protocol, like that for the $n = 3$ protocol, is that players trim pieces of the cake to create ties, rendering them indifferent among these pieces. When $n > 3$, however, one needs to start the trimming and choosing process—leading to an envy-free partial allocation—with *more* pieces than there are players.

As an informal illustration of how to achieve an envy-free *partial* allocation, suppose there are four people. Have Player 1 cut the cake into 5 equal pieces. Player 2 then trims 2 pieces, creating a 3-way tie for largest. Player 3 then trims 1 piece, creating a 2-way tie for largest. The players now choose in the order: Player 4, Player 3, Player 2, Player 1, with the middle two players required to take a piece they trimmed if one is available. Clearly, each player thinks his piece is at least tied for largest. The burden of our demonstration of the n -person envy-free protocol is to show that a full allocation of the entire cake can be accomplished in a *finite* number of steps.

For simplicity, we will present only the $n = 4$ version of the envy-free protocol. The extension to arbitrary n is fairly straightforward and left to the reader. In outline form, the protocol goes as follows:

One player (chosen here to be Player 2 for later notational simplicity), cuts the cake into 4 equal pieces, hands these out, and asks if anyone objects. If, say, Player 1 objects, then Players 1 and 2 (alone) go through several steps which yield six sets (the Y s and Z s in Step 7 below) to be used as a starting partition (in place of the five equal pieces) for the kind of trim-and-choose sequence among all four players that we illustrated two paragraphs earlier. This trim-and-choose sequence is repeated again and again until we arrive at a partial allocation in which Player 1 has an irrevocable advantage over Player 2 (the “aside” after Step 15 below). From this point on, we never have to worry about Player 1’s objecting because of envy for Player 2. Repeating this at most once for each pair of players results in an envy-free allocation of the entire cake after finitely many steps.

Envy-Free Protocol for Arbitrary n
(the $n = 4$ version)
(1992)

- Step 1. Player 2 cuts the cake into 4 pieces (that he considers to be the same size), keeps one piece, and hands one piece to each player.
- Step 2. Each of the other three players is asked whether or not he objects to this allocation. (A player objects iff he envies some other player.)
- Step 3. If no one objects, then each keeps the piece he was given in Step 1, and we are done.
- Step 4. Otherwise, we choose the smallest i so that Player i objected. For notational simplicity, assume $i = 1$. Player 1 now chooses a piece originally given to some other player (whom he envied) and calls that piece A . The piece originally given to Player 1 is called B .
- Aside. Once we have A and B , the other two pieces in the allocation from Step 1 are reassembled. That part of the cake will be allocated later. Note that Player 1 thinks A is larger than B . Player 2 thinks A and B are the same size.
- Step 5. Player 1 now names a positive integer $r \geq 10$ (chosen so that, for any partition of A into r sets, Player 1 will prefer A , even with the 7 smallest—according to Player 1—pieces in the partition of A removed, to B).
- Aside. Player 1 can easily choose such an r . That is, the union of the 7 smallest pieces is certainly no larger than 7 times the average size of all r pieces. Hence, Player 1 simply chooses r large enough so that $7\mu(A)/r < \mu(A) - \mu(B)$, where μ is his measure.
- Step 6. Player 2 now partitions A into exactly r sets (that he considers to be the same size), and does the same to B .
- Step 7. Player 1 chooses (the smallest) 3 sets from the partition of B and names these Z_1, Z_2, Z_3 . He also chooses either (the largest) 3 sets from the partition of A (if he thinks these are all strictly larger than all the Z s), and trims at most 2 of these (to the size of the smallest among the three), or he partitions (the largest) one of the sets in the partition of A into 3 pieces (that he considers to be the same size). In either case, he names these Y_1, Y_2, Y_3 .
- Aside. Player 1’s strategy in Step 7 guarantees that he will think all three Y s are the same size, and each strictly larger than all three Z s. This is

- true even if he chooses the second option.² Player 2 thinks all three Z s are the same size, and each is at least as large as all three Y s.
- Step 8. Player 3 takes the collection of 6 pieces, and either passes (if he thinks there already is at least a 2-way tie for largest), or trims (the largest) one of these (to the size of the next largest), thus creating at least a 2-way tie for largest).
- Step 9. Players 4, 3, 2, and 1, in that order, choose a piece from among the 6 Y s and Z s as modified in Step 8 (that they consider to be largest or tied for largest), with Player 3 required to take the piece he trimmed if it is available. Player 2 must choose Z_1 , Z_2 , or Z_3 . Player 1 must choose Y_1 , Y_2 , or Y_3 .
- Aside. This yields a partition $\{X_1, X_2, X_3, X_4, L_1\}$ of the cake such that $\{X_1, X_2, X_3, X_4\}$ is an envy-free partial allocation, and L_1 is the leftover piece. Moreover, Player 1 thinks his piece X_1 is *strictly larger*—say by ε —than Player 2's piece X_2 .
- Step 10. Player 1 names a positive integer s (chosen so that $[4\mu_1(L_1)/5]^s < \varepsilon$, where μ_1 is Player 1's measure).
- Aside. The integer s specifies how many times the players will iterate the basic trim-and-choose sequence to follow. Notice that if the rules were instead to allow the iterations to continue until Player 1 said “stop” (which he could strategically do at the point at which he thinks the leftover crumb is smaller than the advantage he has over Player 2), then there is no guarantee that a strategically misguided Player 1 would not keep the game going forever.
- Step 11. Player 1 cuts L_1 into 5 pieces (that he considers to be the same size).
- Step 12. Player 2 takes the collection of 5 pieces, selects (the largest) 3 pieces, and trims (the largest) 2 or fewer of these (to the size of the smallest, thereby creating at least a 3-way tie for largest).
- Step 13. Player 3 takes the collection of 5 pieces, perhaps trimmed in step 12, selects (the largest) two, and trims, if he wants to, (the largest) one of these (to the size of the smallest, thus creating at least a 2-way tie for largest).
- Step 14. Players 4, 3, 2, and 1, in that order, choose a piece (that they consider

²The proof runs as follows: We are assuming that both A and B have been partitioned into r pieces, and that B is not only smaller than A but smaller even than A with the smallest 7 pieces of A 's partition removed. Arrange the sets in both partitions from largest to smallest as A_1, A_2, \dots, A_r and B_1, B_2, \dots, B_r . Let μ denote Player 1's measure, and suppose, for contradiction, that both of the following hold:

1. $\mu(B_{r-2}) \geq \mu(A_3)$, which holds if A_1, A_2 , and A_3 are *not* all strictly larger than B_{r-2}, B_{r-1} , and B_r .

2. $\mu(B_{r-2}) \geq \mu(A_1)/3$, which holds if A_1 *cannot* be partitioned into 3 sets all larger than B_{r-2}, B_{r-1} , and B_r .

It follows from 1 that:

3. $\mu(B_7 \cup \dots \cup B_{r-3}) \geq \mu(A_3 \cup \dots \cup A_{r-7})$, since there are $r - 9$ sets in each union, and the smallest one of the B s is at least as large as the largest one of the A s.

It follows from 2 that:

4. $\mu([B_1 \cup B_2 \cup B_3] \cup [B_4 \cup B_5 \cup B_6]) \geq \mu(A_1 \cup A_2)$, since each of the blocks of 3 B s is larger than each of the A s.

But 3 and 4 clearly demonstrate that:

5. $\mu(B) \geq \mu(A_1 \cup \dots \cup A_{r-7})$.

This is the desired contradiction since the set on the right is A with the smallest 7 pieces of its partition removed.

- to be largest or tied for largest), with Players 3 and 2 required to take a piece they trimmed if one is available.
- Step 15. Steps 11–14 are repeated $s - 1$ more times, with each application of these four steps applied to the leftover piece from the preceding application.
- Aside. This yields a partition $\{X'_1, X'_2, X'_3, X'_4, L_2\}$ of the cake such that $\{X'_1, X'_2, X'_3, X'_4\}$ is an envy-free partial allocation, and such that Player 1 thinks that X'_1 is larger than $X'_2 \cup L_2$. We now declare that Player 1 has an *irrevocable advantage* over Player 2, and we begin creation of a subset of $\{1, 2, 3, 4\} \times \{1, 2, 3, 4\}$, which we call “ \mathcal{IA} ” for “irrevocable advantage,” by putting $(1, 2) \in \mathcal{IA}$.
- Step 16. Player 2 cuts L_2 into 12 pieces (that he considers to be the same size).
- Step 17. Each of the other players declares himself to be of type A (if he agrees all the pieces are the same size), or type D (if he disagrees). Player 2 is declared to be of type A .
- Step 18. If $D \times A \subset \mathcal{IA}$, then we give the 12 pieces to the players in A , with each of them receiving the same number of pieces. In this case, we are done.
- Step 19. Otherwise, we choose the lexicographically least pair (i, j) from $D \times A$ that is *not* in \mathcal{IA} , and we return to Step 4 with Player i in the role of Player 1, Player j in the role of Player 2, and L_2 in place of the cake.
- Step 20. Steps 5–18 are repeated.
- Aside. Each time we pass through Step 15, we add an ordered pair to \mathcal{IA} . Notice that since $D \times A \subset \{1, 2, 3, 4\} \times \{1, 2, 3, 4\}$, and $\mathcal{IA} \subset \{1, 2, 3, 4\} \times \{1, 2, 3, 4\}$, we must have $D \times A \subset \mathcal{IA}$ after at most 16 iterations. At this point, we conclude at Step 18 with an envy-free division of the entire cake.

This ends the protocol.

There is an important way (pointed out to us by several people) that the envy-free protocol for even $n = 4$ differs from the envy-free protocol for $n = 3$: For $n = 3$, the number of cuts needed is at most 5, *regardless* of what the measures are. For $n = 4$, the number of cuts needed can be made arbitrarily large by a suitable choice of the four measures (although the moving-knife solution [8] for the four-person problem gives a bounded number of cuts). This raises:

Question 1. Is there a *bounded* envy-free protocol for $n = 4$ or $n > 4$?

There is another slightly more subtle (and perhaps related) way in which the envy-free protocol differs from the others: The three earlier ones also work in the context of what are called “CD preference relations” in [2]. (A CD preference relation is a complete, reflexive, transitive binary relation that satisfies a partitioning postulate, a trimming postulate, and a weak additivity postulate.) The envy-free protocol, on the other hand, requires what is called an “Archimedian CD preference relation” in [2]. The main result in [2] is the fact that a CD preference relation is induced by a finitely additive measure in the obvious way iff it is Archimedian. This raises:

Question 2. Is there an envy-free protocol for $n = 4$ or $n > 4$ that works in the context of *non-Archimedian* preference relations?

It turns out that techniques similar to those used in the n -person envy-free protocol can also be used to solve the “chores” problem [15], wherein each player

wants to minimize the amount of cake he or she receives. This and related questions (e.g., the Pareto-optimality of allocations) are discussed in [6].

REFERENCES

1. A. K. Austin, Sharing a cake, *Mathematical Gazette* 66, no. 437 (1982), 212–215.
2. J. Barbanel and A. Taylor, Preference relations and measures in the context of fair division, *Proceedings of the American Mathematical Society* (to appear).
3. S. Bennett, et al., *Fair Divisions: Getting Your Fair Share*, HIMAP [High School Mathematics and Its Applications] Module 9 (Teachers' Manual), 1987.
4. M. Berliant, K. Dunz, and W. Thomson, On the fair division of a heterogeneous commodity, *Journal of Mathematical Economics* 21 (1992), 235–240.
5. S. J. Brams and A. D. Taylor, A note on envy-free cake division, *Journal of Combinatorial Theory A* (to appear).
6. S. J. Brams and A. D. Taylor, *Fair Division: Procedures for Allocating Divisible and Indivisible Goods* (to appear).
7. S. J. Brams, A. D. Taylor, and W. S. Zwicker, Old and new moving-knife schemes, *Mathematical Intelligencer* (to appear).
8. S. J. Brams, A. D. Taylor, and W. S. Zwicker, A moving-knife solution to the four-person envy-free cake division problem, preprint.
9. P. J. Campbell, Mathematics, *1995 Encyclopedia Britannica Yearbook of Science and the Future* (1994), 379–383.
10. L. E. Dubins and E. H. Spanier, How to cut a cake fairly, *American Mathematical Monthly* 68 (1961), 1–17.
11. S. Even and A. Paz, A note on cake-cutting, *Discrete Applied Mathematics* 7 (1984), 285–296.
12. D. Gale, Mathematical entertainments, *Mathematical Intelligencer* 15, no. 1 (1993), 48–52.
13. *For All Practical Purposes: Introduction to Contemporary Mathematics*, 3d ed., W. H. Freeman, New York, 1994, ch. 13.
14. G. Gamow and M. Stern, *Puzzle-Math*, Viking, New York, 1958.
15. M. Gardner, *aha! Insight*, W. H. Freeman and Company, New York (1978), 123–124.
16. T. P. Hill, Stochastic inequalities, *IMS Lecture Notes* 22 (1993), 116–132.
17. T. P. Hill, Fair-division problems, preprint.
18. B. Knaster, Sur le probleme du partage pragmatique de H. Steinhaus, *Annales de la Societ  Polonaise de Mathematique* 19 (1946), 228–230.
19. S. X. Levmore and E. E. Cook, *Super Strategies for Puzzles and Games*, Doubleday, Garden City, NY (1981), 47–53.
20. A. A. Liapounoff, On completely additive vector functions, *Izv. Akad. Nauk SSSR* (1940), 465–478.
21. J. Neyman, Un theoreme d'existence, *C. R. Acad. Sci. Paris* 222 (1946), 843–845.
22. D. Olivastro, Preferred shares, *The Sciences*, March/April (1992), 52–54.
23. D. Olivastro, "Object Lessons" (Solutions and Sequelae), *The Sciences*, July/August (1992), 55.
24. K. Rebman, How to get (at least) a fair share of the cake, in *Mathematical Plums*, Ross Honsberger, editor, Mathematical Association of America (1979), 22–37.
25. H. Steinhaus, The problem of fair division, *Econometrica* 16 (1948), 101–104.
26. H. Steinhaus, Sur la division pragmatique, *Econometrica* (supplement) 17 (1949), 315–319.
27. W. Stromquist, How to cut a cake fairly, *American Mathematical Monthly* 87, no. 8 (1980), 640–644. Addendum, vol. 88, no. 8 (1981), 613–614.
28. W. Stromquist and D. R. Woodall, Sets on which several measures agree, *J. Math. Anal. Appl.* 108 (1985), 241–248.
29. W. A. Webb, An algorithm for a stronger fair division problem, preprint.
30. D. Weller, Fair division of a measurable space, *Journal of Mathematical Economics* 14, no. 1 (1985), 5–17.
31. D. R. Woodall, Dividing a cake fairly, *J. Math. Anal. Appl.* 78, no. 1 (1980), 233–247.
32. D. R. Woodall, A note on the cake-division problem, *Journal of Comb. Theory (A)* 42 (1986), 300–301.

Department of Politics
New York University
New York, NY 10003

Department of Mathematics
Union College
Schenectady, NY 12308

The Mathematics Portfolio

Mary L. Crowley and Ken Dunn

Portfolios are for art students, right? Well . . . not exclusively. We are proposing the portfolio as an academic artifact for the mathematics major. This article describes our vision of what a mathematics portfolio could be, how to organize the portfolio, items which might be included in a portfolio, and portfolio evaluation. In developing these ideas we have considered several portfolio audiences: the student, the departmental advisor, the department, and, to some extent, future associates such as graduate school departments and employers.

THE PORTFOLIO CONTENTS. As with the artist's portfolio, the mathematics portfolio should include a collection of the student's best mathematical work. The variety and breadth of the items which are available depends, of course, on the mathematical "experiences" the student has encountered and on the reasons for assembling the portfolio. There is a place for the traditional products of a mathematics program—assignments and tests. When, however, a wide range of assessment activities are used, a richer profile of the student's talents can be displayed. This can be achieved through "non-traditional" assignments such as journal writings, book reviews, student presentations (captured on audio or video tape), group projects, computer based activities, and open-ended investigations.

As we envision it, however, the portfolio should be more than a display case for outstanding work. We want the portfolio, for example, to chronicle each student's mathematical career, rather like a faculty member's curriculum vitae. Thus, we suggest that a list of the mathematics courses the student has taken be included. The portfolio also provides the opportunity to guarantee that students are exposed to activities which are valued, and which might not be included in any of their course requirements, e.g. writing about mathematics, critiquing books or articles which are mathematical in nature, or reflecting on their own mathematical expectations and progress. Each department will no doubt have special interests which they too could address through the portfolio.

Figure 1 presents a sample suggestion list for a portfolio. Examples from a range of mathematical areas (e.g., item 7), spanning several semesters (e.g., item 4), and reflecting a variety of mathematical experiences are included. Items 3 and 7 are drawn from "traditional" assessment tasks. Items 1 and 2 are biographical in nature. Items 4 and 8 ask students to reflect on their mathematical expectations and success, while items 5 and 6 provide students the opportunity to analyze both mathematical ideas and presentations. Written responses, not just numerical solutions, are required in several items (1, 4, 5, 6, 7, 8). With item 9, the student is given the responsibility of selecting, without direction, something which reflects his or her mathematical prowess. Not all of these items would necessarily be included in a portfolio and, of course, other items might reflect your program more accurately.

1. A mathematical autobiography.
2. An annotated list of courses taken.
3. The student's honours project (if there is one).
4. A journal covering each semester of the junior and senior year. This could begin with a description of what the student hopes to get out of the semester, include mathematical high and low points of the semester, and end with a follow-up as to what was accomplished in the semester.
5. A critique of one (or more) textbooks.
6. Responses to, or reviews of, articles or books that the students are asked to read over the period of their program.
7. The best test or assignment from each of the "core" courses that the students must take as part of their program. A rationale for each choice should be provided.
8. A description of the mathematical insight which most excited the students over their mathematical career and an explanation why. This might be a theorem, some connections relating distinct mathematical disciplines, the use of mathematics in a modelling process, a conjecture, etc.
9. A mathematical item, of their choice. This is the student's chance to put something in the portfolio which they have generated, of which they are proud, and which has not otherwise been requested. A rationale for why the item has been included should accompany the item. It might be a problem which the student struggled with before solving. It might be something the student "discovered". It might be an instance where they applied their mathematical knowledge in another discipline.

Figure 1. Examples of Portfolio Categories

RATIONALE. The primary goal for asking students to *assemble* the portfolio is to encourage them to reflect on their mathematical flexibility and growth, and to help them focus on their mathematical interests. Selecting the items for inclusion requires students to review the work they have completed, to think about its mathematical value, and to observe how they have matured mathematically. This is particularly the case when the portfolio is accumulated over several semesters, or even several years.

The portfolio is also a useful tool for the student's advisor. From a record keeping point of view, for example, a roster of classes can assist in discussion about future course selection. (Indeed, for anyone interested in what mathematics courses have been completed, it is helpful to have all the courses listed together, rather than embedded in the sessional transcripts.) The remaining portfolio materials inform the advisor about the student's concerns and serve as a departure point for discussions about mathematical interests in general, as well as more practical topics such as career choices.

Once completed, the portfolio can be an item of value and of interest to several constituents. For the students, it is a tangible record of their progress and success... more than the transcripts and tattered notebooks we retain from our undergraduate careers. Since the portfolio provides direct and concrete evidence of a student's abilities and way of thinking, graduate schools, and perhaps potential employers, might also be interested in it. Indeed, the well-crafted portfolio can demonstrate the student's arsenal of problem solving tools, ability to organize and communicate mathematically, writing skills, etc.

The Department will also find it of value to look at the variety and depth of the materials represented in each year's portfolios. If a collection of portfolios is reviewed, it can be a constructive way of informing faculty about the departmental

expectations of students, what material is being covered, and what assessment techniques are being used. We are not implying that portfolios be used to “check” on the teaching or assessment techniques of faculty members. We are suggesting, however, that, in aggregate, portfolios may give a glimpse of what is valued, and valuable, in the major program. For many of these same reasons, potential students might also find the portfolios of interest.

ORGANIZING TO USE THE PORTFOLIO. We feel that it is essential that the Department, at least those teaching courses from which portfolio materials might be drawn, participate in deciding the categories for inclusion in the portfolio and the time period over which items will be collected. Minimally, this collaboration informs all faculty of what is expected of majors. More importantly, however, it also gives each faculty member “ownership” in, and an understanding of, the project. At the same time, the discussions which accompany such joint decision making should also contribute to reviewing the major program as a whole, and, in particular, lead to discussions about goals and assessment techniques.

In our institution, every mathematics major draws up a plan of study with an advisor. At that time, or if you have a forum for meeting all your majors together, the portfolio can be introduced and explained. Topics such as the rationale for the portfolio, the items for inclusion, the dates for completion, and the presentation format should be discussed. The introductory process is completed with the student and the “department” signing an agreement which specifies deadlines and lists the categories from which the portfolio items should be chosen.

Finally, the student is, of course, responsible for assembling the portfolio. The advisor, however, oversees this, checking each semester that appropriate, even if only initial or tentative, contributions have been “deposited”. By the student’s last term, all the material for the portfolio should be selected and assembled. A table of contents should then be added as it will greatly facilitate the use (and evaluation) of the portfolio.

EVALUATION. Rather than assign a mark to the portfolio, we suggest that the completed portfolio be a requirement for a recommendation for graduation. If students see the value of the activity, and the importance put in it by the Department, they will want to do a good job. When, however, the assigning of a grade is deemed necessary, perhaps evaluation themes could be identified, e.g. the diversity of problem solving strategies reflected in the work. Students would be informed about the grading criteria at the time the portfolio was introduced and this would be included in the agreement between the department and the student. The student could then select items for inclusion accordingly.

As mentioned earlier, a collection of portfolios can serve an important role in *program* evaluation. Information gleaned from studying a sampling of portfolios can inform a department about the range of evaluation techniques being used, about the level of success students are achieving, about the nature of the students’ experiences, about the level of expectations of faculty and students, and about the attainment of educational goals. Information obtained in this way can provide a unique and previously untapped perspective.

CONCLUSION. A well considered mathematics portfolio provides faculty and students with information about themselves and the program. In deciding what should be included, departments and instructors are forced to articulate what they value and what their goals are for their majors. Analysing the portfolios provides

information about how successfully these goals are being met. On a more individual level, departmental advisors can use the portfolio to inform themselves about student progress and interest. And for the student, generating the portfolio is an exercise in self-evaluation, which results in a concrete overview (i.e. the portfolio) of his or her undergraduate mathematical experience.

We see lots of potential for the portfolio. A modified version, for example, can easily be implemented in a one semester course. Or, portfolios can be used across disciplines. Indeed, we have become so intrigued with the idea that we are proposing its use in a new first year program, the Science Foundation Year, which is being introduced on our campus. The goals of this program are to promote the study of the sciences from an integrated perspective, to address timely and lively issues in the field, and to encourage students to reflect on their own thought processes. An end-of-the-year portfolio, drawing from all the courses the students take, and “extras”, will be required of those enrolled in the program. The items to be included will support the goals of the program. By its very nature, the portfolio provides a framework for coordinating activities and a forum for communication.

*School of Education and Department of Mathematics,
Statistics, and Computing Science
Dalhousie University
Halifax, Nova Scotia B3H 3J5
mcrowley@ac.dal.ca
dunn@cs.dal.ca*

PICTURE PUZZLE
(from the collection of Paul Halmos)



Mr. MATLAB a quarter of a century ago.
(see page 56.)

Derivative Polynomials For Tangent and Secant

Michael E. Hoffman

1. INTRODUCTION. Sometimes problems naturally occur in pairs, and it's best to tackle both at the same time. For instance, consider the problem of finding the n th derivative of $\tan x$. It's not hard to see that there are polynomials P_n of degree $n + 1$ for $n = 0, 1, \dots$ so that

$$\frac{d^n}{dx^n} \tan x = P_n(\tan x).$$

This problem has a natural companion: find the n th derivative of $\sec x$. Here there are polynomials Q_n of degree n so that

$$\frac{d^n}{dx^n} \sec x = Q_n(\tan x) \sec x.$$

The P_n and Q_n are different sequences of polynomials, but they are evidently related. The numbers $P_n(0)/n!$ and $Q_n(0)/n!$ are the coefficients of x^n in the Maclaurin series for $\tan x$ and $\sec x$ respectively, and their computation is a classical problem.

Here's another problem: for positive integer n and $0 < a < 1$, what is the improper integral

$$\int_{-\infty}^{\infty} \frac{x^n e^{ax}}{e^x - 1} dx?$$

It has a natural companion problem where the denominator is replaced by $e^x + 1$. As we shall see, this turns out to be essentially the same pair of problems as considered in the previous paragraph.

This paper has two main parts. First, in §2 and §3 we obtain the polynomials P_n and Q_n as instances of 'derivative polynomials' associated with functions f such that $f'(x)$ is a polynomial function of $f(x)$. Then in §4 we apply this theory to the computation of improper integrals and infinite series, followed by concluding remarks in §5.

2. DERIVATIVE POLYNOMIALS. Suppose f is a function whose derivative is a polynomial in f , i.e. $f'(x) = P(f(x))$ for some polynomial function P . Then all the higher derivatives of f are also polynomials in f , so we have a sequence of polynomials P_n defined by

$$f^{(n)}(x) = P_n(f(x)), \quad n \geq 0.$$

In fact, the polynomials P_n are determined by the conditions

$$P_0(u) = u, \quad P_{n+1}(u) = P'_n(u)P(u) \quad \text{for } n \geq 1. \quad (1)$$

If we form the generating function

$$F(u, t) = \sum_{n=0}^{\infty} \frac{t^n}{n!} P_n(u),$$

then the equations (1) are equivalent to

$$F(u, 0) = u, \quad F_t(u, t) = P(u)F_u(u, t). \quad (2)$$

Also, f has a ‘companion’ function $g = \exp(\int f(x) dx)$ satisfying $g'(x) = f(x)g(x)$, and there are polynomials Q_n so that

$$g^{(n)}(x) = Q_n(f(x))g(x), \quad n \geq 0.$$

It is easy to see that the Q_n are determined by the conditions

$$Q_0(u) = 1, \quad Q_{n+1}(u) = uQ_n(u) + Q'_n(u)P(u) \quad \text{for } n \geq 0. \quad (3)$$

If we define a second generating function $G(u, t)$ by replacing P_n in the definition of $F(u, t)$ with Q_n , then (3) is equivalent to

$$G(u, 0) = 1, \quad G_t(u, t) = P(u)G_u(u, t) + uG(u, t). \quad (4)$$

Our first result gives explicit formulas for the generating functions F and G .

Theorem 2.1. *For functions f and g as above, the corresponding generating functions for their derivative polynomials are given by*

$$F(u, t) = f(f^{-1}(u) + t) \quad \text{and} \quad G(u, t) = \frac{g(f^{-1}(u) + t)}{g(f^{-1}(u))}.$$

Proof: Let $u = f(x)$. If $F(u, t) = f(x + t)$ we have $F(u, 0) = f(x) = u$, and

$$P(u)F_u(u, t) = \frac{du}{dx} f'(x + t) \frac{dx}{du} = f'(x + t) = F_t(u, t),$$

so $F(u, t) = f(x + t)$ satisfies (2) above. This establishes the first formula; the proof of the second formula (using (4)) is similar.

In view of (2) and (4) above, the generating functions F and G are determined by P alone. Indeed $y = f(x)$ is a solution of the differential equation $y'(x) = P(y)$, so P determines f up to a constant (i.e., $f(x)$ can be replaced by $f(x + c)$). Here are two examples.

Example 1. Let $f(x) = kx$, so that $P(u) = k$. The generating function $F(u, t)$ is just $u + kt$. The companion of f is $g(x) = e^{kx^2/2}$, and from the theorem $G(u, t)$ is $\exp(tu + kt^2/2)$. Expand this out to get

$$Q_n(u) = \sum_{2i \leq n} \binom{n}{2i} (2i - 1)(2i - 3) \cdots 1 k^i u^{n-2i}.$$

In the case $k = -1$, the Q_n are (one variant of) the Hermite polynomials.

Example 2. The main example of this paper, of course, is the case $f(x) = \tan x$, $g(x) = \sec x$ (i.e. $P(u) = u^2 + 1$). Here the generating functions are

$$F(u, t) = \frac{u + \tan t}{1 - u \tan t} = \frac{\sin t + u \cos t}{\cos t - u \sin t}$$

and (cf. [10])

$$G(u, t) = \frac{\cos(\tan^{-1} u)}{\cos(\tan^{-1} u)\cos t - \sin(\tan^{-1} u)\sin t} = \frac{1}{\cos t - u \sin t}.$$

Until now the P_n and Q_n have been treated in parallel, but separately. The next result brings them together.

Theorem 2.2. *The generating functions F and G satisfy (and are determined by) the conditions*

$$F(u, 0) = u, \quad G(u, 0) = 1, \quad F_t = P(F), \quad \text{and} \quad G_t = FG.$$

Proof: Using 2.1, we have

$$F_t(u, t) = f'(f^{-1}(u) + t) = P(f(f^{-1}(u) + t)) = P(F),$$

and the second equation is similar.

This result is useful for obtaining recurrences. For instance, in Example 1 above the equation $G_t = FG$ leads to the recurrence

$$Q_{n+1}(u) = P_0(u)Q'_n(u) + nP_1(u)Q_{n-1}(u) = uQ'_n(u) + nkQ_{n-1}(u)$$

since $P_k(u) = 0$ for $k \geq 2$. Compare this with (3) to get $Q'_n(u) = nQ_{n-1}(u)$. It follows that $y(x) = Q_n(x)$ is a solution of the differential equation $ky'' + xy' - ny = 0$. (Of course, this is Hermite's equation if $k = -1$.)

In Example 2, Theorem 2.2 leads to the pair of recurrences

$$\begin{aligned} P_{n+1}(u) &= \sum_{i=0}^n \binom{n}{i} P_i(u) P_{n-i}(u) + \delta_{0n} \quad \text{and} \\ Q_{n+1}(u) &= \sum_{i=0}^n \binom{n}{i} P_i(u) Q_{n-i}(u). \end{aligned} \tag{5}$$

3. COMPUTING PARTICULAR VALUES. Henceforth we specialize to the case $P(u) = u^2 + 1$ (Note this applies equally to $f(x) = \tan x$, $g(x) = \sec x$ and to $f(x) = -\cot x$, $g(x) = \csc x$). In this section we show how to find particular values of P_n and Q_n without computing the polynomials themselves.

Of all the values of the polynomials P_n and Q_n , those at zero are of the most interest: as noted in the introduction, they give the coefficients of the Maclaurin series for tangent and secant. These numbers can be computed as follows. From (1) and (3) it is evident that P_n is an odd function if n is even, and Q_n is odd for n odd. Then using (5):

$$\begin{aligned} P_1(0) &= 1, \quad P_3(0) = \binom{2}{1} P_1(0)^2 = 2, \\ P_5(0) &= \binom{4}{1} P_1(0) P_3(0) + \binom{4}{2} P_3(0) P_1(0) = 16; \end{aligned}$$

and thus

$$\begin{aligned} Q_0(0) &= 1, \quad Q_2(0) = \binom{1}{0} Q_0(0) P_1(0) = 1, \\ Q_4(0) &= \binom{3}{0} Q_0(0) P_3(0) + \binom{3}{2} Q_2(0) P_1(0) = 5, \\ Q_6(0) &= \binom{5}{0} Q_0(0) P_5(0) + \binom{5}{2} Q_2(0) P_3(0) + \binom{5}{4} Q_4(0) P_1(0) = 61. \end{aligned}$$

Other values can be obtained from these via the following functional equation.

Theorem 3.1. If $u \neq 0$, then

$$P_n(u) = 2^n \left[P_n \left(\frac{u^2 - 1}{2u} \right) + \frac{u^2 + 1}{2u} Q_n \left(\frac{u^2 - 1}{2u} \right) \right].$$

Proof: It is enough to show

$$F(u, t) = F \left(\frac{u^2 - 1}{2u}, 2t \right) + \frac{u^2 + 1}{2u} G \left(\frac{u^2 - 1}{2u}, 2t \right)$$

for $u \neq 0$. Let $u = -\cot x$ and apply 2.1 to get $F(u, t) = -\cot(x + t)$ and $\csc xG(u, t) = \csc(x + t)$. From the half-angle formula for tangent,

$$\begin{aligned} F(u, t) &= -\cot(2x + 2t) - \csc(2x + 2t) \\ &= F(-\cot 2x, 2t) - \csc 2xG(-\cot 2x, 2t), \end{aligned}$$

from which the result follows.

Putting $u = 1$ in Theorem 3.1, we have

$$P_n(1) = 2^n (P_n(0) + Q_n(0)) = \begin{cases} 2^n Q_n(0), & n \text{ even,} \\ 2^n P_n(0), & n \text{ odd.} \end{cases}$$

To get the $Q_n(1)$ we need the following recurrence, which complements (5).

Theorem 3.2. For $n \geq 0$,

$$(u^2 + 1) \sum_{i=0}^n \binom{n}{i} Q_i(u) Q_{n-i}(u) = P_{n+1}(u).$$

Proof: It suffices to prove the identity $(u^2 + 1)G(u, t)^2 = F_t(u, t)$, whose right-hand side is $P(F) = F(u, t)^2 + 1$ by Theorem 2.2. But by 2.1,

$$\begin{aligned} F(u, t)^2 + 1 &= \tan^2(\tan^{-1}(u) + t) + 1 = \sec^2(\tan^{-1}(u) + t) \\ &= \sec^2(\tan^{-1}(u))G(u, t)^2. \end{aligned}$$

Given the $P_n(1)$, the $Q_n(1)$ can be obtained recursively by setting $u = 1$ in Theorem 3.2. At this point the reader may find it instructive to compute $Q_4(1)$.

Remark. There are more efficient ways to compute the numbers $P_n(0)$ and $Q_n(0)$ than that outlined here: see [3] and [9]. For the $Q_n(1)$ see [8].

4. IMPROPER INTEGRALS AND SERIES. We now use the polynomials P_n and Q_n to express some improper integrals and infinite series, starting with the former.

Theorem 4.1. Let $0 < a < 1$. Then for integer $n \geq 0$,

$$\int_{-\infty}^{\infty} \frac{x^n e^{ax}}{e^x + 1} dx = \pi^{n+1} \csc a \pi Q_n(-\cot a \pi).$$

Proof: The only pole of the meromorphic function $e^{az}/(e^z + 1)$ inside the rectangle with vertices $-R, R, R + 2\pi i$, and $-R + 2\pi i$ is at πi , where it has residue $-e^{\pi ia}$. Integrate it around this rectangle and take limits as $R \rightarrow \infty$ to obtain the result for $n = 0$. Then differentiate n times with respect to a , noting that the n th derivative of $\csc x$ is $\csc x Q_n(-\cot x)$.

If the denominator of the integrand is replaced by $e^x - 1$, we have to modify the numerator in the case $n = 0$ to get a convergent integral.

Theorem 4.2. *Let $0 < a < 1$. Then*

$$(a) \quad \int_{-\infty}^{\infty} \frac{x^n e^{ax}}{e^x - 1} dx = \pi^{n+1} P_n(-\cot a\pi) \quad \text{for integer } n \geq 1, \text{ and}$$

$$(b) \quad \int_{-\infty}^{\infty} \frac{e^{ax} - e^{(1-a)x}}{e^x - 1} dx = -2\pi \cot a\pi.$$

Proof: To prove (a), consider f defined by

$$f(z) = \frac{ze^{az}}{e^z - 1}, \quad z \neq 0; \quad f(0) = 1.$$

Then f is analytic inside the rectangle with vertices $-R, R, R + \pi i$, and $-R - \pi i$, so we can apply Cauchy's theorem and let $R \rightarrow \infty$ to get

$$\int_{-\infty}^{\infty} \frac{xe^{ax}}{e^x - 1} dx = -e^{a\pi i} \int_{-\infty}^{\infty} \frac{(x + \pi i)e^{ax}}{e^x + 1} dx.$$

Now use 4.1 to compute the right-hand side, and (a) follows for $n = 1$. Differentiation with respect to a then gives the general case. For (b), note that the integrand has no poles within the rectangle used to prove (a): integrate it around this contour, take limits as $R \rightarrow \infty$, and simplify using 4.1.

Next we pass from integrals to series.

Theorem 4.3. *For real $0 < a < 1$ and integer $n \geq 0$,*

$$(a) \quad \sum_{k=0}^{\infty} \left[\frac{1}{(k+a)^{n+1}} + \frac{(-1)^{n+1}}{(k+1-a)^{n+1}} \right] = \frac{\pi^{n+1}}{n!} P_n(\cot a\pi)$$

and

$$(b) \quad \sum_{k=0}^{\infty} \frac{(-1)^k}{(k+a)^{n+1}} + (-1)^n \sum_{k=0}^{\infty} \frac{(-1)^k}{(k+1-a)^{n+1}} = \frac{\pi^{n+1}}{n!} \csc a\pi Q_n(\cot a\pi).$$

Proof: For $n \geq 1$, let

$$I_n(a) = \int_0^{\infty} \frac{x^n e^{ax}}{e^x - 1} dx.$$

Since

$$\int_{-\infty}^0 \frac{x^n e^{ax}}{e^x - 1} dx = \int_0^{\infty} \frac{(-x)^n e^{-ax}}{e^{-x} - 1} dx = (-1)^{n+1} I_n(1-a)$$

and $P_n(-x) = (-1)^{n+1} P_n(x)$, 4.2(a) can be restated as

$$I_n(1-a) + (-1)^{n+1} I_n(a) = \pi^{n+1} P_n(\cot a\pi).$$

On the other hand,

$$I_n(a) = \int_0^{\infty} \frac{x^n e^{(a-1)x}}{1 - e^{-x}} dx = \sum_{k=0}^{\infty} \int_0^{\infty} x^n e^{-(k+1-a)x} dx = \sum_{k=0}^{\infty} \frac{n!}{(k+1-a)^{n+1}},$$

and so (a) follows for $n \geq 1$. For $n = 0$, note that

$$\int_{-\infty}^{\infty} \frac{e^{ax} - e^{(1-a)x}}{e^x - 1} dx = 2 \int_0^{\infty} \frac{e^{ax} - e^{(1-a)x}}{e^x - 1} dx = 2 \sum_{k=0}^{\infty} \left[\frac{1}{k+1-a} - \frac{1}{k+a} \right]$$

and use 4.2(b). The proof of (b) using 4.1 and the series expansion

$$\int_0^{\infty} \frac{x^n e^{ax}}{e^x + 1} dx = \sum_{k=0}^{\infty} \frac{(-1)^k n!}{(k+1-a)^{n+1}}$$

is entirely analogous (but simpler, since no special argument is required for $n = 0$).

Remark. Equation 4.3(a), for rational a , appears implicitly in §171 of Euler's algebra text [6]. In two earlier papers [4], [5] he relates the left-hand sides of 4.3(a) and 4.3(b) (for rational a) to improper integrals of certain rational functions.

From Theorem 4.3 we can derive a quite useful result on series of powers of reciprocals of integers. We shall need the following definitions. A function $\psi: \mathbf{Z} \rightarrow \mathbf{C}$ is periodic mod q if $\psi(q) = 0$ and $\psi(n+q) = \psi(n)$ for all $n \in \mathbf{Z}$, and alternating mod q if $\psi(q) = 0$ and $\psi(n+q) = -\psi(n)$ for all $n \in \mathbf{Z}$. If ψ is a periodic or alternating function mod q , then it is even if $\psi(q-j) = \psi(j)$ for all $1 \leq j \leq q$, and odd if $\psi(q-j) = -\psi(j)$ for all $1 \leq j \leq q$.

Theorem 4.4. *If ψ is a periodic function mod q , then*

$$(a) \quad \sum_{j=1}^{\infty} \frac{\psi(j)}{j^{n+1}} = \frac{\pi^{n+1}}{2q^{n+1}n!} \sum_{p=1}^{q-1} \psi(p) P_n \left(\cot \frac{p\pi}{q} \right)$$

where n is even and ψ is odd, or n is odd and ψ even. If ψ is an alternating function mod q , then

$$(b) \quad \sum_{j=1}^{\infty} \frac{\psi(j)}{j^{n+1}} = \frac{\pi^{n+1}}{2q^{n+1}n!} \sum_{p=1}^{q-1} \psi(p) \csc \frac{p\pi}{q} Q_n \left(\cot \frac{p\pi}{q} \right)$$

when n and ψ are either both even or both odd.

Proof: As the proofs of (a) and (b) are similar, we give only the former. Set $a = p/q$ in 4.3(a) and multiply both sides by $\psi(p)/q^{n+1}$; then sum over p to get

$$\sum_{p=1}^{q-1} \sum_{k=0}^{\infty} \left[\frac{\psi(p)}{(qk+p)^{n+1}} + \frac{(-1)^{n+1} \psi(p)}{(qk+q-p)^{n+1}} \right] = \frac{\pi^{n+1}}{q^{n+1}n!} \sum_{p=1}^{q-1} \psi(p) P_n \left(\cot \frac{p\pi}{q} \right).$$

If n and ψ have opposite parity, $(-1)^{n+1} \psi(p) = \psi(q-p)$ and the left-hand side is

$$\sum_{p=1}^{q-1} \sum_{k=0}^{\infty} \frac{\psi(p)}{(qk+p)^{n+1}} + \sum_{p=1}^{q-1} \sum_{k=0}^{\infty} \frac{\psi(q-p)}{(qk+q-p)^{n+1}} = 2 \sum_{j=1}^{\infty} \frac{\psi(j)}{j^{n+1}}.$$

Here are some examples: for the first two, it's helpful to recall §3.

Example 1. Since the periodic function mod 2 with $\psi(1) = 1$ is even, we have

$$1 + \frac{1}{3^{n+1}} + \frac{1}{5^{n+1}} + \frac{1}{7^{n+1}} + \cdots = \frac{\pi^{n+1}}{2^{n+2}n!} P_n(0),$$

for n odd, from which it follows that

$$1 + \frac{1}{2^{n+1}} + \frac{1}{3^{n+1}} + \frac{1}{4^{n+1}} + \cdots = \frac{\pi^{n+1} P_n(0)}{2(2^{n+1} - 1)n!} \quad (6)$$

for such n . Similarly, using the alternating function mod 2 with $\psi(1) = 1$, we obtain

$$1 - \frac{1}{3^{n+1}} + \frac{1}{5^{n+1}} - \frac{1}{7^{n+1}} + \cdots = \frac{\pi^{n+1}}{2^{n+2}n!} Q_n(0) \quad (7)$$

for even n .

Example 2. By use of alternating functions mod 4, we have

$$1 + \frac{1}{3^{n+1}} - \frac{1}{5^{n+1}} - \frac{1}{7^{n+1}} + \frac{1}{9^{n+1}} + \frac{1}{11^{n+1}} - \cdots = \frac{\sqrt{2} \pi^{n+1}}{4^{n+1}n!} Q_n(1)$$

if n is even, and

$$1 - \frac{1}{3^{n+1}} - \frac{1}{5^{n+1}} + \frac{1}{7^{n+1}} + \frac{1}{9^{n+1}} - \frac{1}{11^{n+1}} - \cdots = \frac{\sqrt{2} \pi^{n+1}}{4^{n+1}n!} Q_n(1)$$

if n is odd. Cf. [7], [8].

Example 3. Dirichlet L -series are defined as follows. Let χ be a homomorphism from the units of $\mathbf{Z}/k\mathbf{Z}$ to the nonzero complex numbers. Extend χ to a function on \mathbf{Z} (called a Dirichlet character mod k) by defining $\chi(n)$ to be $\chi(k\mathbf{Z} + n)$ if $k\mathbf{Z} + n$ is a unit of $\mathbf{Z}/k\mathbf{Z}$ (i.e. $(n, k) = 1$), and 0 otherwise (so χ is periodic mod k as defined above). For complex s , the Dirichlet L -series corresponding to χ is

$$L(s, \chi) = \sum_{n=1}^{\infty} \frac{\chi(n)}{n^s} = \prod_{p \text{ prime}} \left(1 - \frac{\chi(p)}{p^s} \right)^{-1}.$$

If χ is a Dirichlet character mod k , then either $\chi(k-1) = 1$ or $\chi(k-1) = -1$: χ is called even in the first case and odd in the second (which agrees with our previous terminology). By Theorem 4.4, for χ a Dirichlet character mod k and $n \geq 1$ an integer,

$$L(n, \chi) = \frac{\pi^n}{2k^n(n-1)!} \sum_{j=1}^{k-1} \chi(j) P_{n-1} \left(\cot \frac{j\pi}{k} \right)$$

if χ and n are both even or both odd. Cf. [1], [11].

5. CONCLUSION. We have emphasized throughout the symmetry between the P_n and the Q_n . Mostly they are developed in parallel, but in the recurrences and functional equation of §3 they intertwine in an essential way.

It is more usual to state the Maclaurin series for tangent and secant, and the closed form for the series in (6) and (7), in terms of Bernoulli and Euler numbers (see e.g. [2]). The knowledgeable reader may be wondering how the P_n and Q_n are related to the Bernoulli and Euler polynomials. It turns out that the rational values of the Bernoulli and Euler polynomials can be expressed in terms of the P_n and Q_n , but the relation is not a simple one. For the questions considered here, P_n and Q_n seem more natural.

REFERENCES

1. Tom M. Apostol, Dirichlet L -functions and character power sums, *J. Number Theory* 2 (1970).
2. M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, National Bureau of Standards, Washington, D.C., 1964.
3. M. D. Atkinson, How to compute the series expansions of $\sec x$ and $\tan x$, *Am. Math. Monthly* 93 (1986), 387–389.
4. L. Euler, De inventione integralium si post integrationem variabili quantitati determinatus valor tribuatur, *Misc. Berolin.* 7 (1743), 129–171, Reprinted in *Opera Omnia*, ser. I, vol. 17, B. G. Tuebner, Berlin, 1914, pp. 35–69.
5. L. Euler, De summis serierum reciprocarum ex potestatibus numerorum naturalium ortarum dissertatio altera in qua eadem summationes ex fonte maxime diverso derivantur, *Misc. Berolin.* 7 (1743), 172–192, Reprinted in *Opera Omnia*, ser. I, vol. 14, B. G. Tuebner, Berlin, 1924, pp. 138–155.
6. L. Euler, *Introductio in Analysin Infinitorum*, Lausanne, 1748, Reprinted in *Opera Omnia*, ser. I, vol. 8, B. G. Tuebner, Berlin, 1922.
7. J. W. L. Glashier, On the Bernoullian function, *Quarterly J. of Pure and Appl. Math.* 29 (1898), 1–168.
8. J. W. L. Glashier, On the coefficients in the expansions of $\cos x/\cos 2x$ and $\sin x/\cos 2x$, *Quarterly J. of Pure and Appl. Math.* 45 (1914), 187–222.
9. Donald E. Knuth and Thomas J. Backhotz, Computation of Tangent, Euler, and Bernoulli Numbers, *Mathematics of Computation* 21 (1967), 663–688.
10. C. Krishnamachary and M. Bhimasena Rao, On a table for calculating Eulerian numbers based on a new method, *Proc. London Math. Soc.* (2)22 (1923), 73–80.
11. I. J. Zucker and M. M. Robertson, Some properties of Dirichlet L -series, *J. Phys. A: Math. Gen.* 9 (1976), 1207–1214.

Department of the Navy
United States Naval Academy
Annapolis, MD 21402-5000
meh@sma.usna.navy.mil

Wandering Thoughts

A century later, Jonathan Edwards would be keeping strict accounts of his spiritual life in a journal he begins . . . when he is nineteen. The journal seems to have grown out of a series of resolutions Edwards made, and against which he would sometimes dovetail the record of his actual conduct. . . . He can berate himself for feeling “dull, dry, and dead” on a given day, and although he is aware of the dangers to one’s health from excessive self-mortification, he has enough resoluteness to commit himself to an occasional cold shower of mathematics. “When I am violently beset with temptations, or cannot rid myself of evil thoughts,” he resolves, on July 27, 1723, “to do some Arithmetic, or Geometry, or some other study, which necessarily engages all my thoughts, and unavoidably keeps them from wandering.”

From *A Book of One’s Own* by Thomas Mallon,
pp. 106–107, Ticknor & Fields, New York, 1984

The Law of Large Numbers and $\sqrt{2}$

Thomas M. Liggett and Peter Petersen

Sukru Yuksel, a graduate student in architecture at UCLA, wondered how medieval architects approximated $\sqrt{2}$. This would be important, for example, if they wished to design two square rooms in such a way that the area of the larger room was double the area of the smaller. Issues of status might dictate relations of this sort. In considering various possibilities, he discovered the following approach to this approximation problem. Start with the sequence $a_n = (\sqrt{2})^n = (1, \sqrt{2}, 2, 2\sqrt{2}, 4, 4\sqrt{2}, \dots)$. All the numbers appearing in this sequence are “known” (i.e., are integers) except $\sqrt{2}$ itself. Approximate the “unknown” $\sqrt{2}$ by something, say 1, so that the resulting sequence becomes $b_n = (1, 1, 2, 2, 4, 4, \dots)$. Then apply the following procedure to this sequence: A new collection of sequences $b_n^{(k)}$ is defined by

$$b_n^{(k)} = b_n^{(k-1)} + b_{n+1}^{(k-1)}, \tag{1}$$

with $b_n^{(0)} = b_n$:

1	1	2	2	4	4	8	8
	2	3	4	6	8	12	16
		5	7	10	14	20	28
			12	17	24	34	48
				29	41	58	82
					70	99	140
						169	239

Yuksel observed that the successive ratios $b_1^{(k)}/b_0^{(k)}$ approximate $\sqrt{2}$. For example, the first few values of this ratio are (rounded to six decimal places) 1, 1.5, 1.4, 1.416667, 1.413793, 1.414286, 1.414201, while the value of $\sqrt{2}$ is 1.4142135 He wondered how general this phenomenon was, so he came to one of us for help. Shun-hui Zhu, a Hedrick Assistant Professor at UCLA, was also involved in the ensuing discussion.

The proof that the ratios approach the desired limit in this particular case is not too hard—one simply computes $b_n^{(k)}$ explicitly in terms of b_n (we will do this shortly—see (2) below), evaluates the resulting sum, and passes to the limit. Prompted by Yuksel’s question, however, we were interested in seeing how generally this procedure works. It turns out that the proof of the resulting theorem involves some elementary versions of basic probabilistic ideas—the law of large numbers and a strengthened form of this called large deviations—and thus provides us with an opportunity to see these ideas in action in a relatively simple

This paper is an outgrowth of an undergraduate colloquium given by the second author at UCLA in the Winter of 1993.

context. Note that in the example above, $\sqrt[n]{b_n} \rightarrow \sqrt{2}$, so that the theorem below applies to this example.

Theorem. Suppose that b_n is a positive sequence which satisfies

$$\lim_{n \rightarrow \infty} \sqrt[n]{b_n} = \lambda > 0.$$

Define $b_n^{(k)}$ by (1). Then

$$\lim_{k \rightarrow \infty} \frac{b_1^{(k)}}{b_0^{(k)}} = \lambda.$$

We begin the proof by computing $b_n^{(k)}$ explicitly in terms of b_n . To do so, simply check by induction that

$$b_n^{(k)} = \sum_{j=0}^k \binom{k}{j} b_{n+j}. \quad (2)$$

The induction step uses the well known identity

$$\binom{k}{j} = \binom{k-1}{j} + \binom{k-1}{j-1}.$$

Let $\gamma(n) = b_n \lambda^{-n}$. Using (2) and writing b_j in terms of $\gamma(j)$ and λ gives

$$b_0^{(k)} = \sum_{j=0}^k \binom{k}{j} b_j = \sum_{j=0}^k \binom{k}{j} \gamma(j) \lambda^j$$

and

$$\begin{aligned} b_1^{(k)} &= \sum_{j=0}^k \binom{k}{j} b_{j+1} = \sum_{j=0}^k \binom{k}{j} \gamma(j+1) \lambda^{j+1} = \sum_{j=1}^{k+1} \binom{k}{j-1} \gamma(j) \lambda^j \\ &= \sum_{j=0}^k \frac{j}{k-j+1} \binom{k}{j} \gamma(j) \lambda^j + \gamma(k+1) \lambda^{k+1}. \end{aligned}$$

We will prove our convergence statement by obtaining a lower estimate for $b_0^{(k)}$ and an upper estimate for $|b_1^{(k)} - \lambda b_0^{(k)}|$. In order to do so, it is useful to divide the above sums by $(\lambda + 1)^k$, obtaining

$$\frac{b_0^{(k)}}{(\lambda + 1)^k} = \sum_{j=0}^k \binom{k}{j} \gamma(j) \left(\frac{\lambda}{\lambda + 1} \right)^j \left(\frac{1}{\lambda + 1} \right)^{k-j}$$

and

$$\frac{b_1^{(k)}}{(\lambda + 1)^k} = \sum_{j=0}^k \frac{j}{k-j+1} \binom{k}{j} \gamma(j) \left(\frac{\lambda}{\lambda + 1} \right)^j \left(\frac{1}{\lambda + 1} \right)^{k-j} + \gamma(k+1) \frac{\lambda^{k+1}}{(\lambda + 1)^k}.$$

The key to the proof of the theorem is the observation that these expressions can be interpreted as expected values of certain random variables. In order to do so, recall that if Y is an integer valued random variable and h is a nonnegative function, then the expectation of $h(Y)$ is given by

$$Eh(Y) = \sum_k h(k) P(Y = k).$$

Therefore

$$\frac{b_0^{(k)}}{(\lambda + 1)^k} = E\gamma(X_k) \quad (3)$$

and

$$\frac{b_1^{(k)}}{(\lambda + 1)^k} = E \frac{X_k}{k - X_k + 1} \gamma(X_k) + \lambda \gamma(k + 1) \left(\frac{\lambda}{\lambda + 1} \right)^k, \quad (4)$$

where X_k is a random variable with the binomial distribution with parameters k and $p = \lambda / \lambda + 1$:

$$P(X_k = j) = \binom{k}{j} p^j (1 - p)^{k-j}.$$

This is the distribution of the number of successes in k independent trials with success probability p .

The Law of Large Numbers says that $X_k \approx kp$ in an appropriate sense. This suggests that (3) and (4) are approximately $\gamma(kp)$ and $(p/1 - p)\gamma(kp) = \lambda\gamma(kp)$ respectively, so that the ratio is approximately λ as desired. The precise statement provided by the Weak Law of Large Numbers in this case is that

$$\lim_{k \rightarrow \infty} P \left[\left| \frac{X_k}{k} - p \right| > \epsilon \right] = 0 \quad (5)$$

for every $\epsilon > 0$. (See Theorem 8 of Chapter 4 of [2], for example.) This is not a strong enough statement to justify the above approximations. We need to know how rapidly the probabilities in (5) tend to zero. This information is provided by the most elementary form of a large deviation result.

The theory of large deviations began in the late 1930's with the work of H. Cramér. It was fully developed by Donsker and Varadhan (and others) in the late 1970's and early 1980's. The excellent book [1] tells the whole story. A large deviation result says that if A_k is a sequence of events with probabilities tending to zero as a consequence of a law of large numbers, then these probabilities tend to zero exponentially rapidly in the sense that

$$\lim_{k \rightarrow \infty} \frac{1}{k} \log P(A_k)$$

exists and is in some sense computable. While this conclusion is certainly not always valid, it holds sufficiently often that it is very useful.

In the present special case, it is not hard to obtain the needed exponential bounds on these probabilities without an appeal to the general theory. It is interesting to note, though, that the computations below are the starting point for the development of large deviation theory. Here is the statement we need in the present context:

Proposition. *For every $\epsilon > 0$ there is an $r < 1$ such that*

$$P \left[\left| \frac{X_k}{k} - p \right| > \epsilon \right] \leq 2r^k \quad (6)$$

for all $k \geq 1$.

Proof: Take $\theta > 0$, and write

$$e^{\theta k(p+\epsilon)} P \left[\frac{X_k}{k} - p > \epsilon \right] \leq E e^{\theta X_k} = [e^\theta p + q]^k, \quad (7)$$

($q = 1 - p$). The inequality is a form of Chebyshev's inequality—we simply replace X_k by its minimum value ($k(p + \epsilon)$) on the event

$$\left\{ \frac{X_k}{k} - p > \epsilon \right\}$$

and replace the nonnegative $e^{\theta X_k}$ by 0 on the complementary event. The equality in (7) can be obtained from the definition of expectation and the binomial theorem:

$$Ee^{\theta X_k} = \sum_{j=0}^k e^{\theta j} \binom{k}{j} p^j (1-p)^{k-j} = [e^{\theta p} + q]^k. \quad (8)$$

Rewriting (7) gives

$$P\left[\frac{X_k}{k} - p > \epsilon\right] \leq [pe^{\theta(q-\epsilon)} + qe^{-\theta(p+\epsilon)}]^k.$$

Noting that $pe^{\theta(q-\epsilon)} + qe^{-\theta(p+\epsilon)}$ has value 1 and derivative $-\epsilon$ at $\theta = 0$, it follows that this function is < 1 for small positive θ . Therefore, for each $\epsilon > 0$ there is an $r < 1$ such that

$$P\left[\frac{X_k}{k} - p > \epsilon\right] \leq r^k.$$

The same argument (with $\theta < 0$) shows that there is an $r < 1$ such that

$$P\left[\frac{X_k}{k} - p < -\epsilon\right] \leq r^k.$$

The proposition is obtained by combining these results (and choosing the larger r).

Returning to the proof of the theorem, the assumption on the b_n 's implies that $\sqrt[n]{\gamma(n)} \rightarrow 1$, so that for any $\delta > 0$, there are constants c and C such that

$$ce^{-\delta n} \leq \gamma(n) \leq Ce^{\delta n} \quad (9)$$

for all $n \geq 0$. It follows from (8) and (9) that

$$c[e^{-\delta p} + q]^k \leq E\gamma(X_k) \leq C[e^{\delta p} + q]^k. \quad (10)$$

Recall that we needed a lower estimate for $b_0^{(k)}$ and an upper estimate for $|b_1^{(k)} - \lambda b_0^{(k)}|$. Referring to (3), we see that the first of these is given by (10). For the second, break up the expectation of the expression obtained from (3) and (4) according to whether $|X_k/k - p| \leq \epsilon$ or $> \epsilon$ to obtain (assuming (6), (9) and (10) hold)

$$\begin{aligned} \frac{|b_1^{(k)} - \lambda b_0^{(k)}|}{(\lambda + 1)^k} &\leq E\left|\frac{X_k}{k - X_k + 1} - \frac{p}{1 - p}\right| \gamma(X_k) + \lambda \gamma(k + 1) p^k \\ &\leq \max_{|u - kp| \leq \epsilon k} \left| \frac{u}{k - u + 1} - \frac{p}{1 - p} \right| E\gamma(X_k) + \lambda \gamma(k + 1) p^k \\ &\quad + \max_{0 \leq u \leq k} \left| \frac{u}{k - u + 1} - \frac{p}{1 - p} \right| Ce^{k\delta} P\left[\left|\frac{X_k}{k} - p\right| > \epsilon\right] \quad (11) \\ &\leq \max_{|u - kp| \leq \epsilon k} \left| \frac{u}{k - u + 1} - \frac{p}{1 - p} \right| E\gamma(X_k) + \lambda \gamma(k + 1) p^k \\ &\quad + \max\left(\frac{p}{1 - p}, \left|k - \frac{p}{1 - p}\right|\right) 2r^k Ce^{k\delta}. \end{aligned}$$

For fixed ϵ , choose $r < 1$ so that (6) holds, then choose δ so that

$$\frac{e^{\delta r}}{e^{-\delta}p + q} < 1 \quad \text{and} \quad \frac{p}{e^{-\delta}p + q} < 1,$$

and the corresponding constants so that (9) and (10) hold. Now let $k \rightarrow \infty$. Finally, let $\epsilon \rightarrow 0$ in order to see by combining (3), (10) and (11) that

$$\frac{|b_1^{(k)} - \lambda b_0^{(k)}|}{b_0^{(k)}}$$

tends to zero as required.

We conclude this note with two remarks. First, there is the issue of rates of convergence. In the general context of the theorem, there is nothing that can be said about this. However, in the context of the example with which we started, one can see by direct computation of (2) that the convergence is exponentially rapid. The exponential rate depends on the irrational being approximated; the error in approximating $\lambda = \sqrt{m}$ for an integer m is of order a constant multiple of

$$\left(\frac{\lambda - 1}{\lambda + 1} \right)^k.$$

Secondly, one can imagine higher dimensional versions of this approximation scheme. For example, one could take a doubly indexed array $a_{m,n} = \lambda_1^m \lambda_2^n$, take an approximating array $b_{m,n}$, and replace (1) by

$$b_{m,n}^{(k)} = b_{m,n}^{(k-1)} + b_{m+1,n}^{(k-1)} + b_{m,n+1}^{(k-1)}.$$

The analysis is similar, with the binomial distribution being replaced by the trinomial distribution with parameters k and

$$\frac{\lambda_1}{1 + \lambda_1 + \lambda_2}, \quad \frac{\lambda_2}{1 + \lambda_1 + \lambda_2}, \quad \frac{1}{1 + \lambda_1 + \lambda_2}.$$

Under an analogous assumption on the behavior of the n th roots, the result is that

$$\lim_{k \rightarrow \infty} \frac{b_{1,0}^{(k)}}{b_{0,0}^{(k)}} = \lambda_1$$

and

$$\lim_{k \rightarrow \infty} \frac{b_{0,1}^{(k)}}{b_{0,0}^{(k)}} = \lambda_2.$$

REFERENCES

1. J.-D. Deuschel and D. W. Stroock, *Large Deviations*, Academic Press, 1989.
2. P. G. Hoel, S. C. Port and C. J. Stone, *Introduction to Probability Theory*, Houghton Mifflin, 1971.

Department of Mathematics
University of California
Los Angeles CA 90024
tml@math.ucla.edu
petersen@math.ucla.edu

Exponentiation in Rings

R. H. Redfield[†]

1. EXPONENTIAL RINGS. Adding an element of a group to itself may be viewed as multiplying the element by two. Analogously, multiplying an element of a ring by itself may be viewed as raising it to the second power. Although the theory of groups with a multiplication (i.e., the theory of rings) is very well developed, there seems to be no axiomatized theory of rings with an exponentiation. The purpose of this paper is to suggest a formal setting for such a theory and to determine some of the properties of the resulting structures. Underlying the exposition are the general references [1, 2].

The natural numbers may be used as exponents for all the elements of any commutative ring. For the real numbers, there are more possibilities. For instance, r^e exists for all real numbers e and all *positive* real numbers r . With this in mind, we consider bases and exponents separately for the general case. Specifically, let R be a ring, let B (the bases) be a multiplicative subsemigroup of (R, \cdot) which does not contain 0, and let E (the exponents) be a semiring with unit element 1. (That is, $(E, +)$ is a semigroup, (E, \cdot) is a semigroup with 1, and \cdot distributes over $+$ from both the left and the right.) A binary operation $B \times E \rightarrow B \subseteq R ((b, e) \rightarrow b^e)$ makes (R, B, E) an *exponential ring* if it has the usual properties for exponentiation (viz., for all $b, d \in B$, and $e, k \in E$, $b^e d^e = (bd)^e$, $b^{(ek)} = (b^e)^k$, $b^{e+k} = b^e b^k$, and $b^1 = b$).

There are many familiar examples of exponential rings. If for any ring R , $R^\# = \{r | r \neq 0\}$, and for any partially ordered ring R , $R^\geq = \{r | r \geq 0\}$, and $R^> = \{r | r > 0\}$, then, with respect to the usual exponentiation, $(R, R^\#, \mathbb{Z}^\geq)$ is an exponential ring for any integral domain R , and $(F, F^\#, \mathbb{Z})$ is an exponential field for any field F . As well, $(\mathbb{R}, X, \mathbb{Z})$ and $(\mathbb{C}, X, \mathbb{Z})$, where X is any of $\mathbb{R}^>$, $\mathbb{Q}^\#$, $\mathbb{Q}^>$, are all exponential fields, as are $(\mathbb{R}, \mathbb{R}^>, \mathbb{Q})$, $(\mathbb{C}, \mathbb{R}^>, \mathbb{Q})$, $(\mathbb{R}, \mathbb{R}^>, \mathbb{R})$ and $(\mathbb{C}, \mathbb{R}^>, \mathbb{R})$, and $(\mathbb{Z}_2, \mathbb{Z}_2^\#, \mathbb{Z}_2)$ with exponentiation $1^0 = 1 = 1^1$ is a finite exponential field [3]. Other examples of exponential rings can be constructed by using products or sets of homomorphisms (see §§2 and 3 respectively).

Still more examples can be built from those above by expanding the semiring of exponents. For if (R, B, E) is an exponential ring and S is any semiring with unit element, then $(R, B, E \times S)$ is an exponential ring with respect to the exponentiation $b^{(e,s)} = b^e$. However, from the point of view of the exponential structure,

These ideas arose from a discussion with a student, Gary Rosys, following a class in ring theory. We were discussing exponentiation as a possible third operation and I remarked that when I was a graduate student, one of my professors said that he had once tried to investigate this but without success. Gary wanted to pursue the idea and eventually I suggested the definition given here and went on to direct his senior thesis [3] on the topic. He stated and proved Proposition 2.2; we worked out the example following Proposition 2.2 one afternoon.

there is no real difference between the exponential rings (R, B, E) and $(R, B, E \times S)$. This sort of trivial extension can be eliminated by considering only exponential rings which are reduced in the following sense.

2. REDUCED EXPONENTIAL RINGS. If (R, B, E) is an exponential ring, then the binary relation defined on E by letting $e \mathbf{K} d$ if and only if $b^e = b^d$ for all $b \in B$ is easily seen to be a congruence relation. If $[e] = [d]$ for $e, d \in E$, then $b^e = b^d$ for all $b \in B$ and hence the exponentiation $b^{[e]} = b^e$ is well-defined on B ; it is easy to check that, with respect to this exponentiation, $(R, B, E/\mathbf{K})$ is an exponential ring. An exponential ring (R, B, E) is then *reduced* if \mathbf{K} is the identity congruence on E .

In contrast to the construction described at the end of §1, there is an upper bound on the sets of exponents which are available for making a given ring and set of bases a reduced exponential ring. Specifically, let R be a ring and let B be a multiplicative subsemigroup of R not containing 0. Then $\text{Hom}(B, B)$ is a semiring with unit element with respect to the operations $(e + k)(b) = e(b)k(b)$ and $(ek)(b) = (k \circ e)(b)$, and $(R, B, \text{Hom}(B, B))$ is a reduced exponential ring with respect to the exponentiation $b^e = e(b)$. The semiring $\text{Hom}(B, B)$ is the largest possible semiring of exponents in the following sense.

Proposition 2.1. *Suppose that (R, B, E) is an exponential ring and for $e \in E$, let $T(e): B \rightarrow B$ be the function $T(e)(b) = b^e$. Then T is a homomorphism of $(E, +, \cdot)$ into $(\text{Hom}(B, B), +, \cdot)$ with kernel \mathbf{K} and thus if (R, B, E) is reduced, T is one-to-one.*

Proof: If $e, k \in E$ and $b, d \in B$, then $T(e)(bd) = (bd)^e = b^e d^e = T(e)(b)T(e)(d)$ so that $T(e) \in \text{Hom}(B, B)$. Furthermore, $T(e + k)(b) = b^{e+k} = b^e b^k = T(e)(b)T(k)(b) = (T(e) + T(k))(b)$ and $T(ek)(b) = b^{ek} = (b^e)^k = (T(e) \circ T(k))(b) = (T(e)T(k))(b)$, from which it follows that T is a homomorphism. Since $T(e) = T(d)$ if and only if $b^e = T(e)(b) = T(d)(b) = b^d$ for all b , $T(e) = T(d)$ if and only if $e \mathbf{K} d$, and hence T has kernel \mathbf{K} . ■

We say that (R, B^*, E^*) *extends* (R, B, E) if $B \subseteq B^*$ and there exists a one-to-one homomorphism $\tau_E: E \rightarrow E^*$ such that $\tau_E(1) = 1$ and for all $b \in B$ and $e \in E$, $b^{\tau_E(e)} = b^e$. In the reduced case, Proposition 2.1 allows maximal extensions to be determined as follows. Zorn's Lemma produces an extension (R, B^\wedge, E) such that B^\wedge is a maximal semigroup of bases. If (R, B, E) is reduced and commutative, then a fortiori (R, B^\wedge, E) is reduced and commutative and Zorn's Lemma and Proposition 2.1 together produce at least one reduced commutative extension (R, B^\wedge, E^\wedge) such that E^\wedge is a maximal commutative semiring of $\text{Hom}(B^\wedge, B^\wedge)$. (In the noncommutative case, $E^\wedge = \text{Hom}(B^\wedge, B^\wedge)$.) We claim that (R, B^\wedge, E^\wedge) is maximal. For if (R, B^*, E^*) is a reduced commutative extension of (R, B^\wedge, E^\wedge) , then $(R, B^*, \tau_{E^\wedge} \circ \tau_E(E))$ extends $(R, B^\wedge, \tau_{E^\wedge} \circ \tau_E(E))$ and by the maximality of B^\wedge , $B^* = B^\wedge$. Furthermore, since (R, B^\wedge, E^*) extends (R, B^\wedge, E^\wedge) , $T(E^*) \supseteq T(\tau_{E^\wedge}(E^\wedge)) = E^\wedge$, and hence by maximality of E^\wedge , $T(E^*) = T(\tau_{E^\wedge}(E^\wedge))$. Then since T is one-to-one, $E^* = \tau_{E^\wedge}(E^\wedge)$ and hence τ_{E^\wedge} is an isomorphism. It follows that (R, B^\wedge, E^\wedge) is a maximal extension of (R, B, E) .

Note that in the commutative case, E^\wedge may not be all of $\text{Hom}(B^\wedge, B^\wedge)$. For suppose that F is a field of characteristic zero with noncommutative Galois group G over \mathbb{Q} . Then (F, F^*, \mathbb{Z}) is a reduced exponential field for which F^* is a maximal semigroup of bases. Restricting σ in G to F^* yields a function σ_τ in

$\text{Hom}(F^\#, F^\#)$, and if $\sigma \circ \tau \neq \tau \circ \sigma$ in G , then $\tau_r \sigma_r \neq \sigma_r \tau_r$ in $\text{Hom}(F^\#, F^\#)$. So E^\wedge cannot be all of $\text{Hom}(F^\#, F^\#)$. It is nevertheless sometimes possible to recognize maximal semirings of exponents even in the commutative case.

Proposition 2.2 [3]. *If (R, B, E) is a reduced commutative exponential ring for which there exists $\beta \in B$ such that $\{\beta^e | e \in E\} = B$, then E is a maximal semiring of exponents.*

Proof: Suppose that (R, B, E^*) is a reduced commutative extension of (R, B, E) . If $z \in E^*$, then $\beta^z = \beta^s = \beta^{\tau_E(s)}$ for some $s \in E$. But for all $b \in B$, $b = \beta^e$ for some $e \in E$ and $b^z = \beta^{ez} = \beta^{ze} = \beta^{\tau_E(s)e} = b^{\tau_E(s)}$. And since (R, B, E^*) is reduced, $z = \tau_E(s) \in \tau_E(E)$. Thus $E^* = \tau_E(E)$. ■

For example, let $R = \mathbb{Z}$, $B = \mathbb{Z}^>$, and $E = \mathbb{Z}^{\geq}[x]$, and let $p_1 = 2, p_2 = 3, \dots$ denote the prime numbers in ascending order. For $k \in B$, $k = \prod_{i=1}^{\infty} p_i^{e[i]}$, where the $e[i]$ are nonnegative and eventually zero. For $n \in \mathbb{N}$, let $k^{x^n} = \prod_{i=1}^{\infty} p_i^{e[i+n]}$, and for $k \in B$ and $f(x) = a_0 + a_1x + \dots + a_nx^n \in E$, let $k^{f(x)} = k^{a_0+a_1x+\dots+a_nx^n} = (k^{a_0})(k^x)^{a_1} \dots (k^{x^n})^{a_n}$. It is straightforward to check that (R, B, E) is a reduced commutative exponential ring [3]. But for all $i > 1$, $p_i = 2^{x^{i-1}}$, and hence $B = \{2^{f(x)} | f(x) \in E\}$. So by Proposition 2.2, E is a maximal semiring of exponents.

For another example, suppose that R is a commutative ring, that E is a commutative semiring with unit element, and that $\exp: E \rightarrow R^\#$ is a one-to-one function such that $\exp(x+w) = \exp(x)\exp(w)$. If $B = \exp(E)$, then it is easy to check that (R, B, E) is a commutative exponential ring with respect to the exponentiation $b^x = \exp(\exp^{-1}(b)x)$. If $b^x = b^w$ for all $b \in B$, then in particular $\exp(1)^x = \exp(1)^w$ and hence $x = \exp^{-1}(\exp[\exp^{-1}(\exp(1))x]) = \exp^{-1}(\exp(1)^x) = \exp^{-1}(\exp(1)^w) = w$ so that (R, B, E) is reduced. Furthermore, for all $b \in B$, $b = \exp(1)^{\exp^{-1}(b)}$, and hence $B = \{\exp(1)^x | x \in E\}$. It follows from Proposition 2.2 that E is a maximal semiring of exponents.

Proposition 2.2 also allows certain maximal extensions to be recognized. For instance, consider $(\mathbb{R}, \mathbb{R}^>, \mathbb{R})$. If (\mathbb{R}, B, E) is a reduced commutative extension of $(\mathbb{R}, \mathbb{R}^>, \mathbb{R})$ and $b \in B$, then $b^2 = r \in \mathbb{R}^>$ and hence $b = b^1 = b^{\tau_{\mathbb{R}}(1)} = b^{\tau_{\mathbb{R}}(2)\tau_{\mathbb{R}}(1/2)} = (b^2)^{\tau_{\mathbb{R}}(1/2)} = r^{1/2} \in \mathbb{R}^>$. Then $B = \mathbb{R}^>$ and hence Proposition 2.2 implies that $E = \tau_{\mathbb{R}}(\mathbb{R})$. It follows that $(\mathbb{R}, \mathbb{R}^>, \mathbb{R})$ has no nontrivial reduced commutative extensions.

3. ORDERED EXPONENTIAL RINGS. If (R, B, E) is an exponential ring, then B is similar to the positive cone of a compatible partial order on R in the sense that $BB \subseteq B$. If B is additively closed and cancelative, then B indeed forms such a cone. For by hypothesis, $BB \subseteq B$ and $B + B \subseteq B$. Furthermore, if $x \in B \cap (-B)$, then $0 = x + (-x) \in B$, a contradiction. Thus $B \cap (-B) = \emptyset$ and therefore B is the cone of strictly positive elements of the following compatible order on R : $r < s$ if and only if $s - r \in B$.

Sometimes this partial order on R determines a compatible partial order on E . For if E is a ring and $1 \in B$, then for all $b \in B$, $b^0b^1 = b^1 = 1b^1$, and since B is cancelative, $b^0 = 1$. Thus since $b^{-1}b = b^0 = bb^{-1}$, B is a partially ordered group. But then the order-preserving functions in $\text{Hom}(B, B)$ form the positive cone of a partial order compatible with both $+$ and \cdot , and thus by Proposition 2.1, if (R, B, E) is reduced, E inherits a compatible partial order from $\text{Hom}(B, B)$: $e \leq k$ in E if and only if $c^{k-e} \leq d^{k-e}$ whenever $c \leq d$ in B . However, since $1^{k-e}1^{k-e} = 1^{k-e}$ and B is a group, $1^{k-e} = 1$. Thus, if $1 \leq b$ and $c^{k-e} \leq d^{k-e}$ whenever $c \leq d$,

then $b^e = b^e 1^{k-e} \leq b^e b^{k-e} = b^k$, and if $c \leq d$ and $b^e \leq b^k$ whenever $1 \leq b$, then $c^{k-e} = c^k (c^{-1}d)^e d^{-e} \leq c^k (c^{-1}d)^k d^{-e} = d^{k-e}$. It follows that the order on E may also be described: $e \leq k$ in E if and only if $b^e \leq b^k$ whenever $1 \leq b$ in B .

This alternative description of the order on E also makes sense when E is only a semiring and thus we may use it for a general definition. Specifically, define an *ordered exponential ring* to be an exponential ring (R, B, E) for which R is a partially ordered ring, E is a partially ordered semiring, $1 \in B = R^>$, and if $e \leq k$ in E , then $b^e \leq b^k$ whenever $1 \leq b$ in B .

Examples. The construction above shows that any exponential ring (R, B, E) for which E is a ring, $1 \in B$, and B is additively closed and cancelative, may be turned into an ordered exponential ring. As well, if F is a totally ordered field, $(F, F^>, \mathbb{Z})$ is an ordered exponential field, as is $(\mathbb{R}, \mathbb{R}^>, \mathbb{R})$. For another example, note that $(\mathbb{R} \times \mathbb{R}, \mathbb{R}^> \times \mathbb{R}^>, \mathbb{R} \times \mathbb{R})$ is a commutative exponential ring with respect to pointwise addition and multiplication and the pointwise exponentiation $(r, s)^{(e, k)} = (r^e, s^k)$. Taking $\mathbb{R}^> \times \mathbb{R}^>$ as the strictly positive cone for both the ring and the exponents makes $(\mathbb{R} \times \mathbb{R}, \mathbb{R}^> \times \mathbb{R}^>, \mathbb{R} \times \mathbb{R})$ an ordered exponential ring.

Note that the underlying ring in the last, non-totally ordered, example is not lattice-ordered. This is not unusual. For suppose that (R, B, E) is an ordered commutative exponential ring such that E is a ring and B is cancelative. If $x > 0 < y$ in R , then $x + y \in B$, and since (as we observed above) B is a group in this situation, $(x + y)^{-1} \in B$. Then $z = xy(x + y)^{-1} \in B$ and hence $0 < z$. But since $y < x + y$, $y(x + y)^{-1} < 1$ so that $z < x$ and similarly (since B is commutative) $z < y$. It follows that R is an antilattice in the sense that the greatest lower bound (or least upper bound) of two elements exists if and only if the elements are comparable, and thus that in this case, R is in fact totally ordered whenever it is lattice-ordered.

The example above shows that R may be an antilattice which is not totally ordered. What conditions force R to be totally ordered? In the case where R is a directed integral domain and E is a ring containing $1/2$, then R is totally ordered if and only if squares in R are positive. For if R is totally ordered, clearly squares are positive, and if squares are positive, then in particular $(b - 1)^2 \geq 0$ for any $b \in B$. Since $1/2 \in E$, $(b - 1)^2 = z^2$ for some $z \in B \cup \{0\}$. Then $(b - 1 + z)(b - 1 - z) = 0$, and since R is an integral domain, either $b - 1 = -z \leq 0$ or $b - 1 = z \geq 0$. Then B is totally ordered and thus, since R is directed and $B = R^>$, R is totally ordered.

Order-theoretic interactions between bases and exponents. Requiring that squares be positive involves no restriction on the partial ordering of the exponents. However, for some ordered exponential rings, order-theoretic properties of the bases are related to similar properties of the exponents. For example, recall that a partially ordered group is *archimedean* if $g = 1$ whenever $1 \leq g^n \leq h$ for all positive integers n and some $h > 1$, and suppose that (R, B, E) is a directed reduced exponential ring for which E is a ring and B is an archimedean group. We claim that E must be archimedean as well. Note first that since B is a group and $b^0 b^1 = b^1 = b^1 b^0$ for all $b \in B$, $b^0 = 1$ for all b . Thus if $b^e = 1$ for all b , then $b^e = b^0$ for all b , and since (R, B, E) is reduced, $e = 0$. Hence if $0 < ne \leq k$ for all $n > 0$, then $b^e \neq 1$ for some b . Since R (and hence B) is directed, there exists $d \in B$ such that $1 \leq d$, $b \leq d$ and $b^{-1} \leq d$, i.e., such that $1 \leq d$, $1 \leq db^{-1}$ and $1 \leq db$. Then $1 = d^0 \leq d^e$. If $d^e = 1$, then $1 = (db^{-1})^0 \leq (db^{-1})^e = (b^{-1})^e = (b^e)^{-1}$ and $1 = (db)^0 \leq (db)^e = b^e$ so that $b^e = 1$, a contradiction. If $d^e > 1$, then

since $(d^e)^n = d^0 d^{ne} \leq d^{k-ne} d^{ne} = d^k$ for all $n > 0$, B is not archimedean, also a contradiction. It follows that E must be archimedean.

The opposite relationship sometimes occurs when the exponents are totally ordered. Of course, $(\mathbb{R} \times \mathbb{R}, \mathbb{R}^+ \times \mathbb{R}^+, \mathbb{Z})$ is an ordered exponential ring with respect to the usual exponentiation; so if the order on the exponents is to influence the order on the bases (and hence on the ring), the set of exponents must be as large as possible. With this (and Proposition 2.2) in mind, suppose that (R, B, E) is a directed exponential ring for which there exists $1 \leq \beta \in B$ such that $B = \{\beta^e | e \in E\}$. If $a, c \in B$, then $a = \beta^x$ and $c = \beta^y$, and if E is totally ordered, then either $x \leq y$ or $x \geq y$. But if $x \leq y$, then $a = \beta^x \leq \beta^y = c$; and if $x \geq y$, then $a = \beta^x \geq \beta^y = c$. It follows that B is totally ordered and hence that R is as well.

Note finally that since the only totally ordered archimedean rings with nontrivial multiplications are the subrings of \mathbb{R} [1, p 126], the observations above show that the only directed reduced exponential rings (R, B, E) whose exponents E form a totally ordered ring, whose bases B form an archimedean group, and for which $B = \{\beta^e | e \in E\}$ for some $1 \leq \beta \in B$, are ordered exponential subrings of $(\mathbb{R}, \mathbb{R}^+, \mathbb{R})$.

REFERENCES

1. L. Fuchs, *Partially Ordered Algebraic Systems*, Pergamon Press, Oxford, UK, 1963.
2. N. Jacobson, *Basic Algebra I*, W. H. Freeman and Co., New York, USA, 1985.
3. G. Rosys, An introduction to a formalized exponentiation, Department of Mathematics and Computer Science, Hamilton College, 1991.

Department of Mathematics and Computer Science
Hamilton College
Clinton, New York 13323-1292
rredfiel@itsmail1.hamilton.edu

From the *Chronicle of Higher Education* September 14, 1994

PENN STATE, Erie

Applications are invited for tenure-track assistant professor positions beginning fall 1995. Applicants must have a strong commitment to undergraduate teaching. Ph.D. is required; teaching and postdoctoral experience is desirable. Successful candidates will be expected to develop an externally funded research program that involves undergraduates.

...

MATHEMATICS: We seek someone with interests in field of applied mathematics, numerical analysis, or differential equations. Teaching responsibilities will include abstract algebra and topology.

Submitted by *Paul Schaefer*

Integer Hexahedra Equivalent to Perfect Boxes

Blake E. Peterson and James H. Jordan

1. INTRODUCTION. A *perfect box (cuboid)* would be a rectangular parallelepiped with all edges and all diagonals having integer lengths. Even though great mathematicians have searched for a *perfect box (cuboid)* for many centuries, the existence of a *perfect box* is still an unsolved problem. Euler established an infinite collection of non-similar ‘good tries’ where only the interior diagonals failed to have an integer length. The smallest of Euler’s examples had edges of length 44, 117 and 240. For further background on the problem refer to Vic Klee and Stan Wagon’s recent book [4] or Richard Guy’s book [2].

An *integer hexahedron* is a six faced polyhedron with all edges and all diagonals having integer lengths. Since a *perfect box (cuboid)* would be a special case of an *integer hexahedron* a natural question would be “Are there *integer hexahedra*?”. This has been answered in the affirmative. In 1988 Heiko Harborth and Anfried Kemnitz [3] displayed a very small example, Figure 1, that can be described as adjoining the bases of two congruent tetrahedra when their bases are equilateral triangles that have side length 3 and the other faces are all isosceles triangles with equal sides of length 2. (The interior diagonal is also of length 2). Möller [5] provides a description of numerous other examples which are pyramids with a cyclic integer pentagon as a base and congruent isosceles triangles as the other five

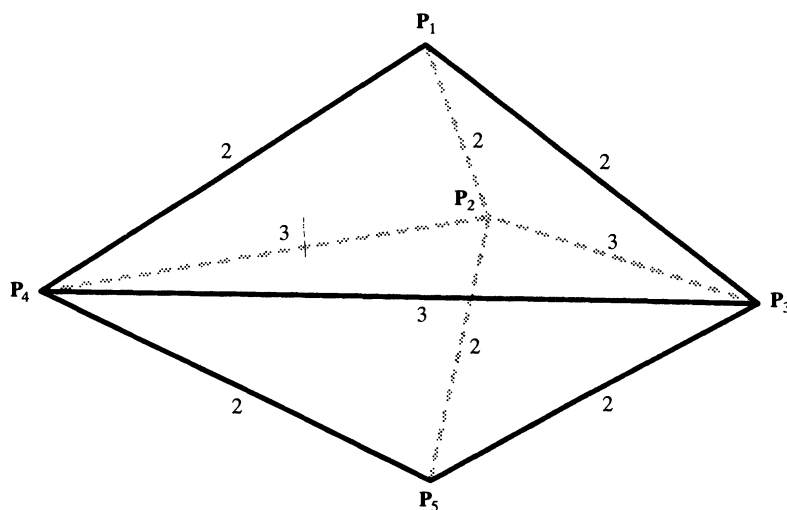
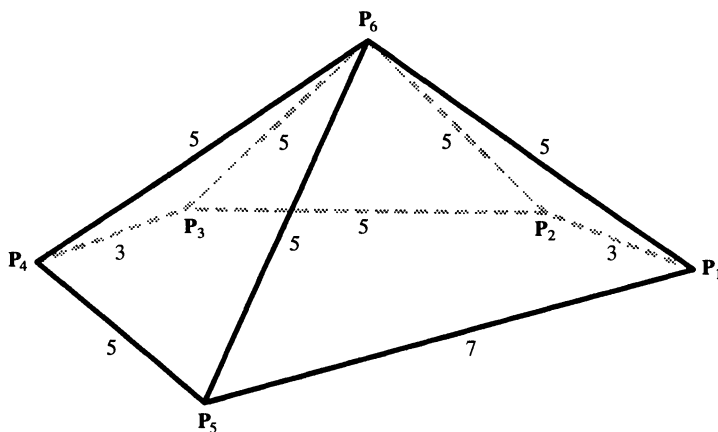
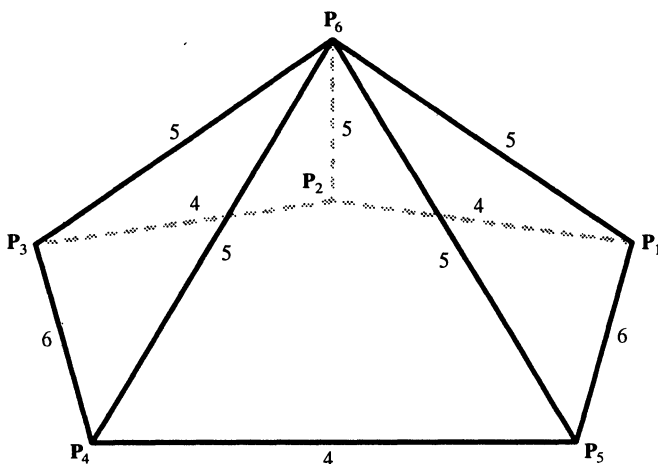


Figure 1. Double Tetrahedron



a.



b.

Figure 2. Small Pentagonal-Based Hexahedron

faces where the triangles' equal sides have length an integer larger than the radius of the circle that circumscribes the cyclic pentagon. The smallest two of these are displayed in Figure 2. The first of the examples has as the base an integer pentagon that is a subset of an integer hexagon which was known by Euler. The second of the examples has as the base an integer pentagon that is displayed in an article by Müller [6]. The other non similar examples involving integer pentagons as bases of pyramids are explicitly described by Möller [5].

The examples of Harborth and Kemnitz or Möller are not combinatorially equivalent to a *perfect box* since they have only five or six vertices as opposed to eight and their faces are triangles or pentagons as opposed to quadrilaterals (rectangles). A question about a combinatorial equivalent to a *perfect box* would be "Are there integer hexahedra with eight vertices that have all six faces quadrilaterals?." It is the purpose of this paper to answer this question in the affirmative by displaying numerous nonsimilar examples.

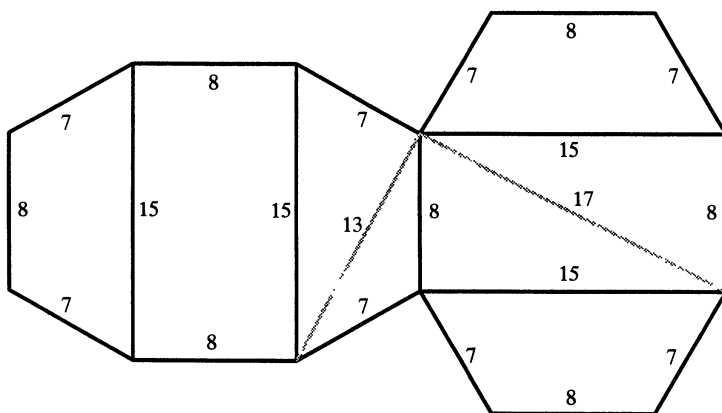


Figure 3. Planar Depiction of Figure 4

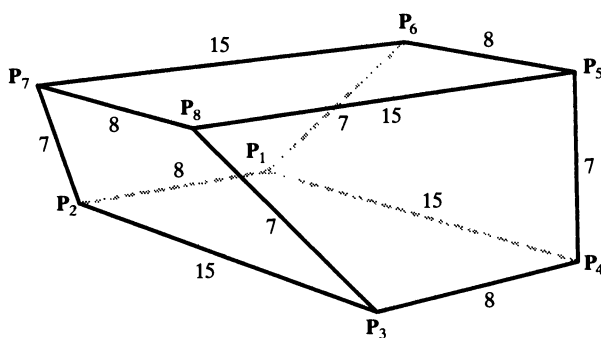


Figure 4. Smallest Example

2. OUR SMALLEST EXAMPLE. Consider two congruent rectangles with sides of length 8 and 15 and diagonals of length 17 and four congruent integer isosceles trapezoids with parallel sides of length 8 and 15, slant edges of length 7, and diagonals of length 13. Place them on a plane in the array of Figure 3. These can be folded together into the *integer hexahedron* of Figure 4 where P_3 , P_4 , P_6 , and P_7 determine the vertices of a planar isosceles trapezoid with parallel sides 8 and 15 and slant sides 13.

The distance between the parallel planes of the rectangles is $7/\sqrt{2}$ and the interior diagonals are all of length 17. A possible eight points in Euclidean Three Space that describes the vertices are: $(\pm 4, \pm 15/2, 7/\sqrt{8})$, $(\pm 15/2, \pm 4, -7/\sqrt{8})$. The volume of the example is approximately 634.39, it has surface area about 934.55, its vertices lie on a sphere of approximate radius of 8.85. The sum of the lengths of all its edges is 120 and the sum of the lengths of all its diagonals is 240.

3. OTHER EXAMPLES. All of the following examples consist of two congruent rectangles and four congruent isosceles trapezoids that can be formed into an *integer hexahedron* in the same manner as the smallest example.

No.	Rectangles		Trapezoids			Interior
	sides	diag	bases	sides	diag	diag
2	120–119	169	120–119	89	149	191
3	88–105	137	88–105	199	221	241
4	280–165	325	280–165	233	317	383
5	288–175	337	288–175	235	325	395
6	280–351	449	280–351	93	327	453
7	352–135	377	352–135	294	366	426
8	160–231	281	160–231	398	442	482
9	288–330	438	288–330	217	377	487
10	360–357	507	360–357	151	389	529
11	280–351	449	280–351	501	591	669
12	432–665	793	432–665	219	579	789
13	936–75	939	936–75	623	677	727
14	560–702	898	560–702	309	699	939
15	432–665	793	432–665	1077	1203	1317
16	520–1302	1402	520–1302	929	1241	1489
17	840–1463	1687	840–1463	601	1261	1679
18	912–1300	1588	912–1300	755	1325	1715
19	864–1330	1586	864–1330	1509	1851	2139
20	728–2310	2422	728–2310	1471	1961	2351

Some points of special interest in these examples should be noted.

- The rectangles of example 2 are very close to being squares and the slant sides are barely tilted. These features make it very close to a *perfect box* although no perfect box could ever have a square face. Are there examples that are even closer?
- The rectangles of example 10 are similar to those of example 2, but the entire figure is not similar. The slant sides of example 10 are tilted a little more than are those of example 2.
- Examples 6 and 11 have the same rectangles as faces and 12 and 15 have the same rectangles as faces but none of these examples is similar to a smaller example.
- Eleven of the twenty examples have rectangles formed of primitive pythagorean triples while the others don't.
- It seems strange that only one example has small positive integers for all lengths and the others require rather large integers.

4. LOCATING EXAMPLES. First recall Ptolemy's Theorem.

Theorem 1 (Ptolemy). *In a cyclic quadrilateral the product of the diagonals is equal to the sum of the products of the opposite sides.*

A proof of Ptolemy's Theorem can be found in Davis [1].

For any Pythagorean Triple, (a, b, c) , there is at least one *integer isosceles trapezoid* whose parallel sides have the lengths a and b , slant sides of length c and diagonals of length d . When two of these rectangles and four of these *integer isosceles trapezoids* are used to build the hexahedron the length of the interior diagonals, designated by f , is the only distance which might not be an integer. This distance f has the property that its square is the area of the rectangle, ab , added to the square of d , the diagonal of the trapezoid. Then it is only a matter of checking the Pythagorean Triple and the associated *integer isosceles trapezoids* to see if f is an integer.

Essentially start with

$$a^2 + b^2 = c^2 \text{ (Pythagoras)}$$

then find e and d such that $d - e < \min(a, b)$ and

$$ab + e^2 = d^2 \text{ (Ptolemy)}$$

and see if there is an integer f such that

$$ab + d^2 = f^2.$$

Note that if f exists then e^2, d^2, f^2 are consecutive terms in an arithmetic progression with common difference ab .

5. REMARKS. Our small example might be the smallest *integer hexahedron* that is combinatorially equivalent to a *perfect box* but we have not proven that.

Because of the many symmetries of the examples and the four sides being congruent, the length of the interior diagonal had to depend only on one rectangular shape and one trapezoidal shape. The interior diagonal of a *perfect box* must depend on the three different rectangles which could be too many restrictions to produce an integer. Our work was more manageable than it would be if we considered different rectangles and trapezoids or even considered other quadrilaterals.

We would like to find a parameter that would give us infinitely many non-similar examples of this type of *integer hexahedron*. Euler's parameter examples all failed to yield a *perfect box* since the one remaining distance was irrational.

REFERENCES

1. Davis, David R. *Modern College Geometry*, Addison-Wesley (1954). pp. 72, 73.
2. Guy, Richard K. *Unsolved Problems in Number Theory*, Springer Verlag, 1981.
3. Harborth, H. and Kemnitz, A. *Diameters of Integer Point Sets*, Colloquia Mathematica Societatis Janos Bolyai 48. Intuitive Geometry, Siofok, (1985) pp. 255–266.
4. Klee, V. and Wagon, S. *Old and New Unsolved Problems in Plane Geometry and Number Theory*, MAA Dolciani Mathematical Exposition No. 11, 1991.
5. Möller, Meinhard *Dissertation* (unpublished), Technischen Universität Carolo-Wilhelmina Zu Braunschweig 1990.
6. Müller, A. Auf einen Kreis liegende Punktmengen ganzzahliger Entfernungen, *Elemente der Mathematik* 8, (1953) pp. 37, 38.

Department of Mathematics
Oregon State University
Corvallis, OR 97331-4605

Department of Mathematics
Washington State University
Pullman, WA 99164-3113

I tell them that if they will occupy themselves with
the study of mathematics they will find in it the
best remedy against the lusts of the flesh.

—Thomas Mann (1875–1955)
The Magic Mountain, New York: Alfred A. Knopf, 1927.

NOTES

Edited by: John Duncan

Calculating Normal Probabilities

Richard J. Bagby

For X a random variable with standard normal distribution and $a > 0$, the probability that $0 < X < a$ is

$$P(a) = \frac{1}{\sqrt{2\pi}} \int_0^a e^{-x^2/2} dx.$$

This is probably the best known example of an integral that cannot be evaluated in terms of elementary functions. In this note we develop an elementary approximation to $P(a)$ which arises in a natural manner, is not difficult to use, and gives excellent results over the entire range $0 < a < \infty$. It is, in fact, more accurate than the four-place tables of values of $P(a)$ commonly found in statistics textbooks, and simpler than many of the better known approximations (as described in Johnson and Kotz [1], for example).

Our approximation to $P(a)$ is

$$Q(a) = \frac{1}{2} \left\{ 1 - \frac{1}{30} \left[7e^{-a^2/2} + 16e^{-a^2(2-\sqrt{2})} + \left(7 + \frac{\pi}{4}a^2 \right) e^{-a^2} \right] \right\}^{1/2},$$

a formula resulting from simple variants of well-known techniques for evaluating or approximating integrals.

The first step is to write

$$\begin{aligned} P(a)^2 &= \frac{1}{2\pi} \int_0^a \int_0^a e^{-(x^2+y^2)/2} dy dx \\ &= \frac{1}{\pi} \int_0^a \int_0^x e^{-(x^2+y^2)/2} dy dx \end{aligned}$$

and then use polar coordinates to obtain

$$P(a)^2 = \frac{1}{\pi} \int_0^{\pi/4} \int_0^{a \sec \theta} e^{-r^2/2} r dr d\theta = \frac{1}{4} - \frac{1}{\pi} \int_0^{\pi/4} e^{-(1/2)a^2 \sec^2 \theta} d\theta.$$

This integral is easier to approximate than the original one, chiefly because the integrand is less variable. Indeed, it becomes constant as either $a \rightarrow 0$ or $a \rightarrow \infty$, and just about any quadrature scheme gives the right values for constant functions.

Our formula for $Q(a)$ comes from an obscure but extremely effective approximation to this last integral. It uses the quadrature rule

$$\begin{aligned} \int_a^b f(x) dx &\approx \frac{1}{30} (b-a) \left[7f(a) + 16f\left(\frac{a+b}{2}\right) + 7f(b) \right] \\ &\quad - \frac{1}{60} (b-a)^2 [f'(b) - f'(a)], \end{aligned}$$

which is exact for polynomials of degree five or less. It is especially efficient for our integral, because $f'(0) = 0$ and $f'(\pi/4) = -2a^2f(\pi/4)$, so that only three functional evaluations are actually used instead of the indicated five. Thus for our integrand it is no harder to evaluate than Simpson's rule, but it is noticeably more accurate.

Like Simpson's rule, our quadrature formula results when the technique of Richardson extrapolation is used to eliminate part of the error inherent in the trapezoidal rule; see any numerical analysis text such as [2] for an extensive discussion of this topic. Since neither our formula nor its error estimate are widely known, we outline a derivation of both that can be extended to develop additional quadrature formulas.

To start our derivation, we need a good representation for the error in the trapezoidal rule, one that allows us not just to bound the error but to estimate it accurately when f is a smooth function. This requires no more than repeated integration by parts, with successive antiderivatives chosen to make most of the boundary terms be zero or antisymmetric. We do all this by choosing an even polynomial $K(x)$ such that $K^{(6)} = 1$, $K(h) = 0$, $K'(h) = 0$, and $K^{(3)}(h) = 0$. Then we have

$$\int_{-h}^h f(x) dx = [K^{(5)}(x)f(x) - K^{(4)}(x)f'(x) - K^{(2)}(x)f^{(3)}(x)] \Big|_{-h}^h \\ + \int_{-h}^h K(x)f^{(6)}(x) dx$$

for all $f \in C^6$. The conditions on K require

$$K(x) = \frac{1}{720}(x^6 - 5h^2x^4 + 7h^4x^2 - 3h^6) = \frac{1}{720}(x^2 - h^2)^2(x^2 - 3h^2),$$

so that with a translation we obtain the one-step formula

$$\int_{c-h}^{c+h} f(x) dx = h[f(c+h) + f(c-h)] - \frac{1}{3}h^2[f'(c+h) - f'(c-h)] \\ + \frac{1}{45}h^4[f^{(3)}(c+h) - f^{(3)}(c-h)] \\ + \frac{1}{720} \int_{-h}^h f^{(6)}(c+x)(x^2 - h^2)^2(x^2 - 3h^2) dx.$$

We can now use our knowledge of the error to eliminate much of it. Applying the one-step rule to the left and right halves separately and adding the results yields the two-step formula

$$\int_{c-h}^{c+h} f(x) dx = \frac{1}{2}h[f(c+h) + 2f(c) + f(c-h)] \\ - \frac{1}{12}h^2[f'(c+h) - f'(c-h)] \\ + \frac{1}{720}h^4[f^{(3)}(c+h) - f^{(3)}(c-h)] \\ + \frac{1}{720} \int_{-h/2}^{h/2} \left[f^{(6)}\left(c - \frac{h}{2} + x\right) + f^{(6)}\left(c + \frac{h}{2} + x\right) \right] \\ \times \left(x^2 - \frac{h^2}{4} \right)^2 \left(x^2 - \frac{3h^2}{4} \right) dx.$$

Simpson's rule (with a representation of the error) comes from subtracting one-third of the one-step formula from four-thirds of the two-step one; the second terms on the right are then eliminated. Our quadrature rule comes from subtracting one-fifteenth of the one-step formula from sixteen-fifteenths of the two-step one, thereby eliminating the terms proportional to h^4 . The terms proportional to h^2 are not eliminated; they become part of our quadrature formula.

To subtract the integrals involving sixth derivatives efficiently, rewrite them in the form

$$\int_0^h [f^{(6)}(c+x) + f^{(6)}(c-x)] p(x) dx.$$

Symmetry helps; in the integral over $[-h/2, h/2]$ one simply replaces x by $\pm(x - \frac{1}{2}h)$. This leads eventually to the identity

$$\begin{aligned} \int_{c-h}^{c+h} f(x) dx &= \frac{1}{15}h[7f(c+h) + 16f(c) + 7f(c-h)] \\ &\quad - \frac{1}{15}h^2[f'(c+h) - f'(c-h)] \\ &\quad + \frac{1}{3600} \int_0^h [f^{(6)}(c+x) + f^{(6)}(c-x)] \\ &\quad \times (x-h)^4(5x^2 + 4hx + h^2) dx. \end{aligned}$$

Since the last integral contains a polynomial with constant sign, the mean value theorem for integrals allows us to express it as $\frac{1}{4725}h^7f^{(6)}(\xi)$, with ξ some value between $c-h$ and $c+h$. The quadrature rule we originally stated corresponds to $h = (b-a)/2$ and $c = (a+b)/2$, and the last integral represents the error.

What does all this say about the accuracy of approximating $P(a)$ by $Q(a)$? Our representation of the error is very hard to use effectively when $f(\theta) = e^{-(1/2)a^2 \sec^2 \theta}$. Computing $f^{(6)}(\theta)$ is a routine but lengthy exercise; a systematic method is to call $t = \tan \theta$ so that $f^{(n)}(\theta) = p_n(t)f(\theta)$ with $p_{n+1}(t) = (1+t^2)[p'_n(t) - a^2tp'_n(t)]$. While the individual terms in this expansion are easy to deal with, the alternating sum in which they appear is not. Moreover, for many values of a the function $f^{(6)}(\theta)$ has sign changes on $[0, \pi/4]$, so its weighted averages are often significantly smaller than its extreme values. All this means that simple bounds for $\frac{1}{4725}h^7f^{(6)}(\xi)$ are a good bit larger than the observed error. Consequently, our statements about the accuracy of approximating $P(a)$ by $Q(a)$ are based not on the error formula, but instead on a detailed comparison of computed values of $Q(a)$ with tabulated values of $P(a)$.

We found that the error $Q(a) - P(a)$ changes slowly in a and varies from about -0.00003 near $a = 0.30$ to $+0.00003$ near $a = 1.70$; as expected, it vanishes both as $a \rightarrow 0$ and as $a \rightarrow \infty$. The formula is remarkably accurate near $a = 0$; even the relative error $(Q(a) - P(a))/P(a)$ remains small as $P(a) \rightarrow 0$. We can prove this last statement by using the estimate $e^{-t} = 1 - t + O(t^2)$ for all the exponentials. That leads to

$$Q(a) = \frac{a}{2} \left[\frac{170 - 64\sqrt{2} - \pi}{120} \right]^{1/2} + O(a^3),$$

while $P(a) = \frac{a}{\sqrt{2\pi}} + O(a^3)$ as $a \rightarrow 0$.

Thus for small a , the relative error is theoretically about -0.0003 , although computing $Q(a)$ accurately as $a \rightarrow 0$ requires some care.

Still better approximations to $P(a)$ are available by our general method. Simply using the same quadrature rule separately on $[0, \pi/8]$ and $[\pi/8, \pi/4]$ and adding the results should divide the error by about 64, since for multistep applications the error formula shows the error is $O(h^6)$. Of course, that requires evaluating $f(\theta)$ at five points instead of three, but $f'(\theta)$ is still evaluated only at $\theta = 0$ and $\pi/4$. If we are willing to evaluate $f(\theta)$ at five or more points, we can also use a quadrature rule of higher order. But $Q(a)$ already gives all the accuracy ever needed for elementary statistics classes.

REFERENCES

1. N. Johnson and S. Kotz, *Continuous Univariate Distributions-1*, Houghton-Mifflin, Boston, 1970.
2. D. Kincaid and W. Cheney, *Numerical Analysis*, Brooks/Cole, Pacific Grove, 1990.

Department of Mathematical Sciences
New Mexico State University
Las Cruces, NM 88003
rbagby@nmsu.edu

Constrained Critical Points

Paul Shutler

1. INTRODUCTION. This note addresses the same problem as that posed in the recent paper by Cassell and Rees [1], namely the computation of the index of a constrained critical point. They treat it from an algebraic point of view. In Section 2 we give an analytic treatment of the restricted Hessian which is at the heart of the problem. In Section 3 we show how to calculate this restricted Hessian by working through the first of the two examples given in [1]. In Section 4 we show how the result in [1], on bordered Hessians, fits more naturally into this analytic, rather than algebraic, framework. Specifically, by using the rank theorem of differential analysis, [2, Section II.7], we can pick a coordinate system in which the algebra of [1] is greatly simplified.

Recall the situation. We have an object function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ which we wish to extremise on a subset M given as the zero locus of another function $g: \mathbb{R}^n \rightarrow \mathbb{R}$. (Reference [1] treats the case of arbitrarily many constraint functions g_i . We shall stick to the case of a single constraint function for the sake of clarity, but everything we shall say extends in an obvious way to the general case.) We introduce the Lagrangian $L = f + \lambda g$ and find a point $p \in M$ and a value $\lambda \in \mathbb{R}$ such that $\nabla_{\mathbb{R}^n} L(p) = 0$. We would like M to be not just a subset but also a codimension one submanifold, so that we can differentiate on it. The easiest way to ensure this is to insist that $\nabla_{\mathbb{R}^n} g \neq 0$ and then apply the rank theorem to g . (This is consistent with the Lagrangian approach, since $\nabla_{\mathbb{R}^n} g(p) = 0$ would mean that p was already a critical point of the unconstrained object function.) It then

follows that $\nabla_M(f|_M)(p) = 0$, hence p is a critical point of the constrained object function $f|_M$. What kind of critical point is it?

2. THE RESTRICTED HESSIAN. The answer to this question is contained in the proposition below. Roughly speaking, it says that the Hessian of the constrained object function $f|_M$ is the restriction to the tangent space of M of the Hessian of the Lagrangian L . A Hessian *matrix*, however, is not something which of its nature admits of restriction to a subspace. We must first convert it into a *bilinear form*. To understand why this step is problematic, and to explain the need for all the subscripts in the statement of the proposition, we must first appreciate the difference between these two kinds of objects.

A Hessian matrix $(\partial^2 f / \partial x_i \partial x_j)$ requires a coordinate system $\mathbf{x} = (x_1, \dots, x_n)$ in order to define it. A bilinear form B on a vector space V does not require a basis in order to define it. If C_1 and C_2 are matrix representations of B with respect to two bases of V then we have $C_2 = A^T C_1 A$, where A is the basis change matrix. We say that the two matrices are *congruent*. Thus, a bilinear form is the same as a collection of congruent matrices, one for each basis. So, if the collection of Hessian matrices of a function f at a point p , one for each coordinate system, is to define a bilinear form on the tangent space of the manifold in question, the Hessian matrices must be congruent to one another. In general they are not. To see this, let \mathbf{x}, \mathbf{y} be two coordinate systems. Then a simple exercise in use of the chain rule gives,

$$\frac{\partial^2 f}{\partial y_i \partial y_j} = \sum_{k,l} \left(\frac{\partial x_k}{\partial y_i} \right) \frac{\partial^2 f}{\partial x_k \partial x_l} \left(\frac{\partial x_l}{\partial y_j} \right) + \sum_k \left(\frac{\partial f}{\partial x_k} \right) \frac{\partial^2 x_k}{\partial y_i \partial y_j}.$$

The matrix $(\partial x_i / \partial y_j)$ is playing the rôle of the basis change matrix A , but the second term spoils the congruence relation unless $\nabla f = (\partial f / \partial x_k)$ is zero.

As an example, consider how tempting it is to claim that $H(f|_M) \sim H(f)|_{T_p M \times T_p M}$ where H denotes a Hessian, \sim denotes congruence, and $T_p M$ denotes the tangent space at p to the manifold M . That this is false can be seen by considering the case $f(x, y, z) = z$, $g(x, y, z) = z - x^2 - y^2$, $p = (0, 0, 0)$. Then $H(f|_M)(p) \sim \text{diag}(1, 1)$ while $H(f)(p) = 0$. What has gone wrong, of course, is that $\nabla f(p) \neq 0$. This may seem like an elementary mistake, but in a moment we shall be making essentially the same claim about L , so it is important to be clear about when it is true and when it is not.

In our constrained extremisation problem we should assume that $\nabla_{\mathbb{R}^n} f(p) \neq 0$, otherwise there would be no point in introducing the Lagrangian. But we do have $\nabla_M(f|_M)(p) = 0$, hence we do get a well defined bilinear form on $T_p M$ which we write $B_{M,p}(f|_M)$. Similarly, $\nabla_{\mathbb{R}^n} L(p) = 0$ gives us $B_{\mathbb{R}^n,p}(L)$. To write $B_{\mathbb{R}^n,p}(f)$, however, would be to write nonsense, which explains why the subscripts and restriction signs are so important.

Proposition:

$$B_{M,p}(f|_M) = B_{\mathbb{R}^n,p}(L)|_{T_p M \times T_p M}.$$

Proof: Choose coordinates \mathbf{x} at p such that M is the locus $x_n = 0$ and (x_1, \dots, x_{n-1}) are coordinates on M . That this can be done follows from applying

the rank theorem to our assumption that $\nabla_{\mathbb{R}^n} g \neq 0$. Then observe that

$$B_{M,p}(L|_M) = B_{\mathbb{R}^n,p}(L)|_{T_p M},$$

since both may be represented by the matrix $(\partial^2 L / \partial x_i \partial x_j)$ for $i, j = 1, \dots, (n-1)$. Finally, since $g|_M \equiv 0$, we have $L|_M \equiv f|_M$. This completes the proof.

3. EXAMPLE. As an illustration of how to apply the above proposition we shall treat the first of the two examples given in [1]. Let

$$f = x^3 + y^3 + z^3 \quad g = x^{-1} + y^{-1} + z^{-1} - 1.$$

(Notice that our choice of g differs from that in [1] by -1 so as to make M the zero locus.) The stationary points are then $p_1 = (3, 3, 3)$ with $\lambda = 243$, $p_2 = (1, 1, -1)$ with $\lambda = 3$, and p_3, p_4 symmetrical with p_2 , that is, with the same value of λ but with the minus sign i.e. the other two spots.

At p_1 the Hessian of L in xyz -coordinates is $\text{diag}(36, 36, 36)$ which is positive definite. Its restriction to any subspace is therefore positive definite too, so we are at a minimum point of $f|_M$.

At p_2 the Hessian of L is $\text{diag}(12, 12, -12)$. To construct a basis for $T_{p_2} M$ observe that this space is orthogonal in the Euclidean sense to $\nabla_{\mathbb{R}^n} g(p_2) = (-1, -1, -1)$. A suitable basis is therefore $\{(1, 1, -2), (-1, 1, 0)\}$, but restricted to this basis the Hessian becomes $\text{diag}(-2, 2)$ so we are at a saddle point of $f|_M$.

4. THE BORDERED HESSIAN. We can view the Lagrangian as a function $\hat{L}(\mathbf{x}, \lambda)$ of the Lagrange multiplier λ as well as of the coordinates \mathbf{x} . Since $\partial \hat{L} / \partial \lambda = g$ vanishes on M , $\nabla_{\mathbb{R}^n \oplus \mathbb{R}} \hat{L}(p, \lambda) = 0$ so we obtain a bilinear form $B_{\mathbb{R}^n \oplus \mathbb{R}, (p, \lambda)}(\hat{L})$. The *bordered Hessian* is the Hessian matrix of \hat{L} . We can go one step further in the proof of the above proposition by using the rank theorem to choose coordinates \mathbf{x} such that $g(\mathbf{x}) = x_n$. Then the Hessian matrix of $\hat{L} = f + \lambda x_n$ takes the simple form

$$\begin{pmatrix} \left(\frac{\partial^2 f}{\partial x_i \partial x_j} \right)_{i,j=1}^{n-1} & * & 0 \\ * & \ddots & \vdots \\ * \cdots \cdots * & 0 & 1 \\ 0 \cdots \cdots 0 & 1 & 0 \end{pmatrix},$$

where the asterisks mark the derivatives $\partial^2 f / \partial x_i \partial x_n$, $i = 1, \dots, n-1$, and where we recognise the upper left hand submatrix as the Hessian of the constrained object function $f|_M$. This matrix is similar to the corresponding matrix in [1, Section 2]. The difference is that our choice of coordinates gives us many zero entries, which allows us to use

$$A = \begin{pmatrix} & 0 & 0 \\ \mathbf{I}_{n-1} & \vdots & \vdots \\ & 0 & 0 \\ 0 \cdots \cdots 0 & 1 & 0 \\ - * \cdots - * & 0 & 1 \end{pmatrix},$$

where \mathbf{I}_{n-1} is the $(n-1) \times (n-1)$ identity matrix, in a congruence transformation $A^T()A$ to eliminate the asterisks. Diagonalising the lower right hand two by

two matrix we deduce

$$B_{\mathbb{R}^n \oplus \mathbb{R}, (p, \lambda)}(\hat{L}) = B_{M, p}(f|_M) \oplus \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

that is, the bordered Hessian is congruent to the constrained Hessian summed with a saddle point pair. This is the main result in [1]. Applying it to the above example thus requires the diagonalisation of a four by four matrix which, although not difficult to do, is rather tiresome. It is worth emphasising again at this point the importance of the distinction between Hessian matrices and bilinear forms. It is the condition $\nabla \hat{L}(p, \lambda) = 0$ which allows us to use bilinear forms to bridge the gap between the above rather special coordinates and the coordinates we are given in the example. Otherwise it would not be clear that our four by four matrix, diagonalised or not, had anything to do with the local behaviour of $f|_M$.

5. CONCLUSION. We have presented two methods for discerning the nature of a constrained critical point. We can construct a basis for the tangent space of the submanifold and diagonalise the $(n - 1) \times (n - 1)$ restricted Hessian matrix. Alternatively, we can diagonalise the $(n + 1) \times (n + 1)$ bordered Hessian matrix. Which of these two methods is to be preferred is perhaps a matter of taste, although in the above example the former turned out to be the easier.

REFERENCES

1. C. Hassell and E. Rees, The Index of a Constrained Critical Point, *Amer. Math. Monthly* 100 (1993) 772–778.
2. W. M. Boothby, *An Introduction to Differentiable Manifolds and Riemannian Geometry*, Academic Press Inc., 1986.

*National Institute of Education
Nanyang Technological University
469 Bukit Timah Road
Singapore 1025
shutlerp@nievax.nie.ac.sg*

A Cone Eversion

S. Tabachnikov

In the punctured plane $\mathbb{R}^2 - (0, 0)$ two functions are given: $f_0(x, y) = \sqrt{x^2 + y^2}$ and $f_1(x, y) = -\sqrt{x^2 + y^2}$. Their gradients are the constant radial vector fields (Fig. 1). Certainly, these fields are homotopic as nondegenerate vector fields (that is, they can be included into a continuous one-parameter family of vector fields without zeroes in the punctured plane): just rotate each vector through 180° . Can one perform such a homotopy in the class of nondegenerate *gradient* vector fields? In other words, can one include the functions $f_0(x, y)$ and $f_1(x, y)$ into a

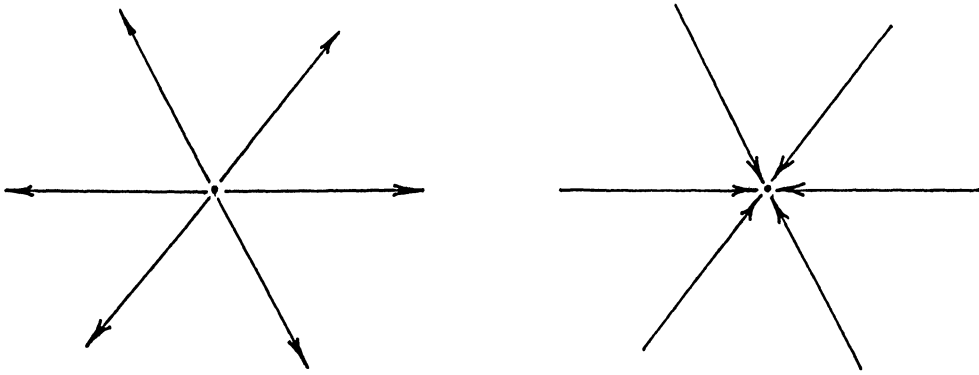


Figure 1.

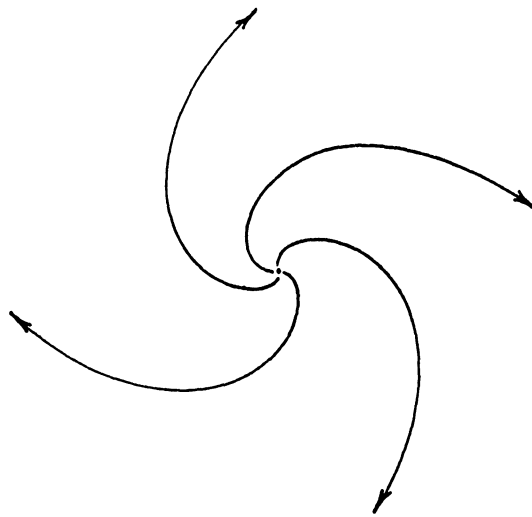


Figure 2.

one-parameter family of smooth functions $f_t(x, y)$ without critical points in the punctured plane, continuously depending upon the parameter t ? The reader is encouraged to make his/her own attempt (warning: rotation of the vectors in the same sense does not work—the field in Fig. 2 is not a gradient!)

Formulate the problem more geometrically. Fix an open annulus in the horizontal x, y -plane, say, the annulus $1 < \sqrt{x^2 + y^2} < 3$. It is diffeomorphic to the punctured plane, so it is enough to solve the problem in the annulus (indeed, if one has a family of functions f_t without critical points in the annulus, composing it with the diffeomorphism yields a desired family of functions in the punctured plane). Consider the function $f_t(x, y)$ as the height function of a surface S_t in space, whose projection onto the horizontal plane is the fixed annulus. The surfaces S_0 and S_1 are cones (Fig. 3); the latter is the “lamp” and the former—the “lump” (in analogy with the “cap” \cap and the “cup” \cup). What one wants to achieve is a deformation of the “lump” S_0 to the “lamp” S_1 , so that each intermediate surface S_t does not have a horizontal tangent plane at any point.



Figure 3.

Here is an example of the deformation in question. The surface S_t is given by the equation

$$z = g_t(\alpha) + 0.25(r - 2)h_t(\alpha)$$

in the cylindrical coordinates (α, r, z) ; here $0 \leq \alpha \leq 2\pi$, $1 < r < 3$ and the “time” parameter t varies from 0 to 4. The functions g and h are:

$$\begin{aligned} g_t &= t \sin \alpha, & h_t &= (1 - t) + t(0.5 + \cos \alpha); & t &\in [0, 1]; \\ g_t &= (2 - t) \sin \alpha + (t - 1) \sin 2\alpha, & h_t &= \cos \alpha + 0.5(2 - t); & t &\in [1, 2]; \\ g_t &= -(t - 2) \sin \alpha + (3 - t) \sin 2\alpha, & h_t &= \cos \alpha - 0.5(t - 2); & t &\in [2, 3]; \\ g_t &= -(4 - t) \sin \alpha, & h_t &= -(t - 3) + (4 - t)(\cos \alpha - 0.5); & t &\in [3, 4]. \end{aligned}$$

The reader may (but probably will not) verify that, for all values of t , the function $z_t(\alpha, r)$ does not have critical points.

One could stop here; but I believe that I owe the reader some explanations. First, the existence of the homotopy in question is a very particular consequence of the Gromov-Phillips theorem and the Gromov h -principle theory (see [G] and [H]). The proofs in this theory are by no means constructive, so explicit constructions are of interest. A famous example is turning a sphere inside out—another consequence of the Gromov theory (more precisely, of its predecessor, the Hirsch-Smale theorem); see, e.g., [Fr] or the movie under preparation at the Minnesota Geometry Center. The problem we are concerned with here was mentioned in [F] (and was given to me by my advisor D. Fuchs some 17 years ago; I believe the present construction is similar to a somewhat obscure one I produced then).

Secondly, I should like to explain how the above formulas came up. Since the original and the terminal functions are linear in r , it is natural to look for the function z_t in the form:

$$z_t(\alpha, r) = g_t(\alpha) + \epsilon(r - 2)h_t(\alpha),$$

where g and h are periodic functions and ϵ is a small parameter to be chosen. The original “lump” surface corresponds to $g_0(\alpha) = 0$ and $h_0(\alpha) = \text{const} > 0$; the terminal “lamp”—to $g_4(\alpha) = 0$ and $h_4(\alpha) = \text{const} < 0$. It might be instructive to think about the surface S_t as a sort of closed rope ladder in space, whose axis is the curve

$$z = g_t(\alpha), \quad 0 \leq \alpha \leq 2\pi, \quad r = 2,$$

and whose rungs are the radial segments

$$z = g_t(\alpha) + \epsilon(r - 2)h_t(\alpha), \quad \alpha = \text{const}, \quad 1 < r < 3$$

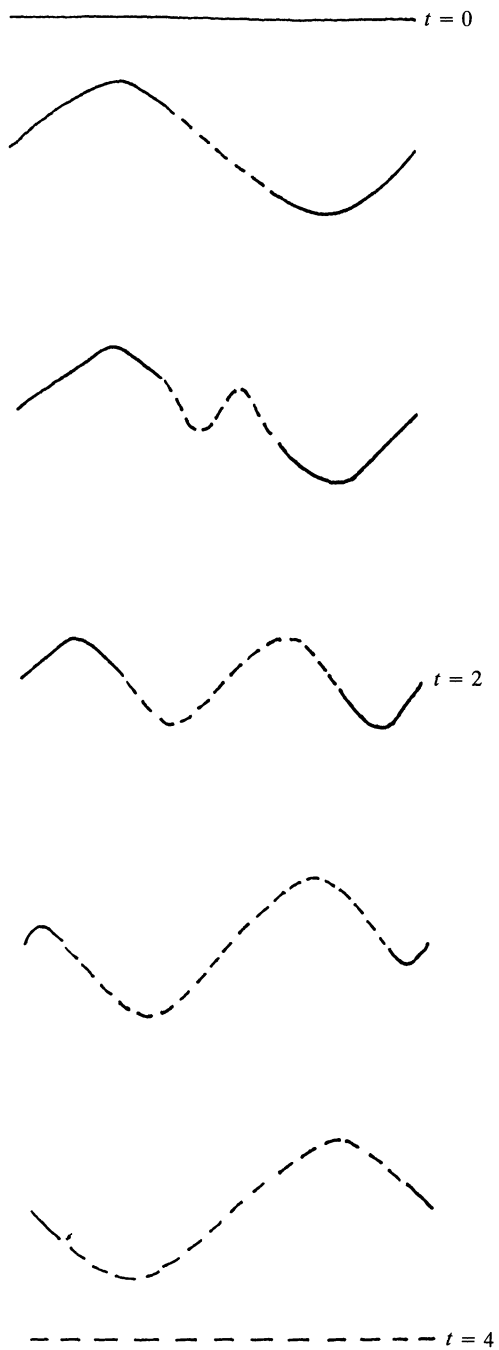


Figure 4.

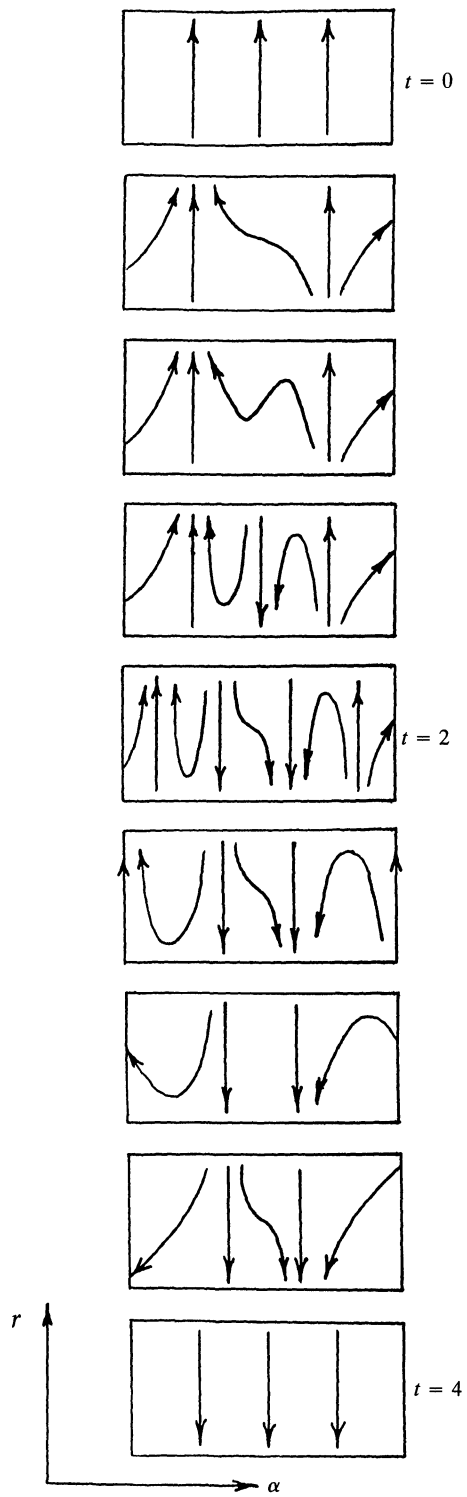


Figure 5.

with the slope of $\epsilon h_t(\alpha)$. So, at the beginning, the axis is a horizontal circle and the slopes of all the rungs are positive. At the end, the axis is again a horizontal circle, but the slopes of the rungs are all negative.

What one wants to avoid in the deformation are any points where the axis and the rungs are simultaneously horizontal. Thus the functions

$$\frac{dg_t(\alpha)}{d\alpha} + \epsilon(r-2)\frac{dh_t(\alpha)}{d\alpha} \quad \text{and} \quad h_t(\alpha)$$

should not have common zeros. If, for some t , the zeros of

$$\frac{dg_t(\alpha)}{d\alpha} \quad \text{and} \quad h_t(\alpha)$$

are disjoint, then so are the zeroes of

$$\frac{dg_t(\alpha)}{d\alpha} + \epsilon(r-2)\frac{dh_t(\alpha)}{d\alpha} \quad \text{and} \quad h_t(\alpha)$$

for a sufficiently small ϵ (use continuity and compactness of the circle). By compactness of the t -interval this ϵ can be chosen uniformly for all $t \in [0, 4]$.

The strategy is clear now. First, change the shape of the axis of the rope ladder (i.e., the graph of $g(\alpha)$) into a non-horizontal curve, after which one can safely change the slope of the rungs (the sign of $h(\alpha)$) from positive to negative on its non-horizontal segments.

The graphs of $g_t(\alpha)$ are sketched in Fig. 4. The graphs are drawn in solid or broken lines; the former means that $h_t(\alpha)$ is positive, and the latter—that it is negative at the corresponding points α . The half-way picture ($t = 2$) is symmetric with respect to the time eversion: $t \rightarrow 2 - t$; from that point on one just repeats the process backwards (should one call this half-way surface S_2 the “limp”?). The reader is encouraged to use his/her favorite software to visualize the “limp” S_2 . Fig. 5 shows the corresponding homotopy of the gradient vector fields, thus answering the original question.

REFERENCES

- [F] D. Fuchs. Cohomology of Infinite-Dimensional Lie Algebras and Characteristic Classes of Foliations. In *Modern Problems in Math.*, v. 10, 179–285 (1978), (in Russian).
- [Fr] G. Francis. *A Topological Picturebook*. Springer, 1987.
- [G] M. Gromov. *Partial Differential Relations*. Springer, 1986.
- [H] A. Haefliger. *Lectures on the Theorem of Gromov*. Springer Lect. Notes Math., v. 209, 128–141 (1971).

Department of Mathematical Sciences
University of Arkansas
301 SCEN
Fayetteville, AR 72701

Answer to Picture Puzzle (p. 22)

Cleve Moler, the principal creator of MATLAB.

THE COMPUTER SCIENCE SAMPLER

Edited by: Catherine C. McGeoch

Approximation Algorithms: Good Solutions to Hard Problems

Ran Libeskind-Hadas

1. INTRODUCTION. Consider a computer network represented by an undirected graph where the vertices represent computer nodes and the edges represent links between the nodes. Since some of the links in the network may become faulty, link testing devices are placed at some of the nodes. A tester at a particular node can test all links incident to that node. Since the testers are expensive, however, we wish to deploy the minimum number of these devices such that every link is incident to at least one node containing a tester. In graph theoretic terms, a *vertex cover* is a subset of the vertices such that every edge is incident to at least one vertex in this set. Our objective then is to find a minimum vertex cover. This is known as the *vertex cover problem*.

The vertex cover problem is one of many computational problems known to be NP-complete (see “Turing Machines and Computational Complexity” in the January 1994 issue of the *Monthly*). NP-complete problems can be solved in a number of steps that grows exponentially in the size of the problem, but no “efficient” algorithms are known for these problems. By “efficient” we mean that the number of steps, or *time*, is bounded by some polynomial in the size of the problem. In fact, not only are no polynomial time algorithms known for NP-complete problems, but the theory of NP-completeness tells us that if a polynomial time algorithm is found for *any* single NP-complete problem, then *all* NP-complete problems are solvable in polynomial time. Theoretical computer scientists generally believe, but have so far been unable to prove, that there do not exist polynomial time algorithms for NP-complete problems.

Let us reconsider the vertex cover problem. A simple algorithm enumerates all possible subsets of the vertices in increasing order of cardinality, and tests each set to see if it is a vertex cover for the graph. This process terminates when the first vertex cover is discovered. In the worst case, this algorithm will terminate at the very last set, since the set of all vertices is certainly a vertex cover. For a graph with n vertices, essentially 2^n subsets must be considered by the algorithm in the worst case. For example, if this algorithm was applied to deploy link testers in a network with 100 nodes and the algorithm was executed on a supercomputer capable of considering 10^{12} subsets per second, the computer would require over 40 billion years to consider all possible subsets! Since the vertex cover problem is NP-complete, it is unlikely that a dramatically faster algorithm can be found for this problem.

What, then, can we do when confronted with an NP-complete problem such as the vertex cover problem? One approach is to use *heuristic algorithms*. These algorithms employ simple rules of thumb and, consequently, tend to be very fast. However, heuristics do not guarantee that an optimal solution, or even anything close to an optimal solution, will be found. A natural heuristic for the vertex problem, for example, begins by selecting a vertex of highest degree (that is, the vertex with the maximum number of edges incident to it), in this way “covering” as many edges as possible. This step is repeated until every edge is covered. Unfortunately, there are many graphs for which this heuristic performs quite poorly. In fact, for any positive value of α , it is possible to construct a graph such that the solution found by the heuristic on this graph is α times larger than the optimal solution [8]! It would certainly be much more desirable to have an algorithm that finds a vertex cover that is guaranteed to be at most some fixed constant times larger than optimal. Such an algorithm is called an *approximation algorithm* and an approximation algorithm that runs in polynomial time is called a *polynomial time approximation algorithm*. The mere existence of polynomial time approximation algorithms is somewhat surprising, since we have no efficient way of determining optimal solutions to NP-complete problems. Using a number of clever techniques, however, researchers have discovered approximation algorithms for many important NP-complete problems.

2. APPROXIMATION ALGORITHMS. We begin by describing a surprisingly simple polynomial time approximation algorithm for the vertex cover problem. Let $G = (V, E)$ denote a given graph. The algorithm comprises the following steps:

1. S is initially the empty set.
2. While edges remain in the graph, select an edge (u, v) arbitrarily. Add u and v to the set S and remove u, v , and all edges incident to these vertices from G .

We claim that when this algorithm terminates, S is a vertex cover for graph G and the cardinality of S is at most twice that of a vertex cover of minimum size.

The first part of this claim is easily established, since at any step the edges remaining in G are exactly those edges that are not yet covered by vertices in S . To verify the second part of this claim, let $E' = \{e_1, \dots, e_k\}$ denote the set of edges selected by the algorithm. By definition, every vertex cover must include at least one of the two endpoints of each of these edges. Observe also that these edges have no vertices in common, since once an edge is selected, both of its endpoints and all incident edges are removed from G . Therefore, every vertex cover, and in particular a minimum vertex cover, must have size at least k . However, the vertex cover constructed by this algorithm has size exactly $2k$ since S consists of both endpoints of each edge in E' . Thus, this algorithm obtains a vertex cover that is at most twice as large as a minimum vertex cover. Finally, it is not difficult to show that this algorithm runs in a number of steps that grows polynomially (in fact linearly) in the number of vertices and edges in the graph.

We have demonstrated a polynomial time approximation algorithm that finds vertex covers that are at most twice as large as optimal. In fact, our analysis is tight: It is not difficult to construct graphs for which this algorithm finds vertex covers that are exactly twice as large as optimal. In general, let A denote an approximation algorithm and let $A(I)$ denote the size of the solution obtained by this algorithm for a particular instance I of the problem. Similarly, let $\text{OPT}(I)$ denote the size of an optimal solution for instance I of the problem. We define the

ratio $R_A(I)$ by

$$R_A(I) = \frac{A(I)}{\text{OPT}(I)}$$

and the *absolute performance ratio* R_A of algorithm A is defined by

$$\inf\{r \mid R_A(I) \leq r, \text{ for all instances } I \text{ of the problem}\}.$$

Is it possible that more sophisticated approximation algorithms for the vertex cover problem achieve absolute performance ratios better than 2? The answer is indeed “yes”, although surprisingly the best algorithm currently known improves this ratio only slightly to $2 - (\log \log n / 2 \log n)$ where n is the number of vertices in the graph [2]. Thus, asymptotically, this algorithm is no better than our simple approximation algorithm. Generalizing the notion of the absolute performance ratio, the *asymptotic performance ratio* R_A^∞ of algorithm A is defined by

$$R_A^\infty = \inf\{r \mid \exists N_0, \text{ s.t. } R_A(I) \leq r, \text{ for all instances } I \text{ of the problem s.t. } \text{OPT}(I) \geq N_0\}.$$

We now turn to another problem, known as the *bin packing problem*, for which approximation algorithms with much better absolute and asymptotic performance ratios are known. In the bin packing problem we are given a finite set of items, each with size between 0 and 1. Our objective is to pack these items into unit capacity bins, minimizing the total number of bins used. More formally, let $I = \{s_1, s_2, \dots, s_n\}, \forall I, s_i \in [0, 1]$ denote the set of items. We wish to partition I into disjoint subsets (bins) B_1, B_2, \dots, B_k such that $\forall i, \sum_{s_j \in B_i} s_j \leq 1$ and k is as small as possible.

Like the vertex cover problem, the bin packing problem is NP-complete. Like the vertex cover problem as well, a very simple polynomial time approximation algorithm for bin packing finds solutions that are at most twice as large as optimal. This approximation algorithm, known as the *first fit algorithm*, operates as follows: Select one item at a time in arbitrary order and place this item in the first bin which can accommodate it.

The ratio of 2 for the first fit algorithm follows from two observations. First, we show that when the algorithm terminates, at most one of the used bins is half full or less. Assume that this is not the case. Then when the algorithm terminates, there are two bins B_i and $B_j, i < j$, that are each at most half full. Then the last item placed in B_j clearly has size at most $\frac{1}{2}$. Since bin B_i has capacity at least $\frac{1}{2}$ throughout execution of the algorithm, the first fit algorithm would have placed this item in B_i rather than in B_j , a contradiction. Now, letting $\text{FF}(I)$ denote the number of bins used by the first fit algorithm on a given problem instance I , this observation implies that

$$\text{FF}(I) < \left\lceil 2 \sum_{s_i \in I} s_i \right\rceil.$$

Our second observation is that the total number of bins used in any solution is at least the sum of the sizes of all the items. In particular, letting $\text{OPT}(I)$ denote the number of bins used in an optimal solution, we have

$$\left\lceil \sum_{s_i \in I} s_i \right\rceil \leq \text{OPT}(I).$$

Combining these two observations, we now have

$$\text{FF}(I) < 2 \cdot \text{OPT}(I)$$

and thus $R_{\text{FF}} < 2$.

The above analysis shows that the absolute performance ratio of the first fit algorithm is less than 2. In fact, more careful analysis shows that for all instances I of the bin packing problem

$$\text{FF}(I) \leq \frac{17}{10} \text{OPT}(I) + 2$$

and that there exist instances I with arbitrarily large values of $\text{OPT}(I)$ such that

$$\text{FF}(I) \geq \frac{17}{10} (\text{OPT}(I) - 1).$$

Therefore, the asymptotic performance ratio of the first fit algorithm R_{FF}^∞ is in fact 1.7. Moreover, a minor modification of the first fit algorithm achieves an even better performance ratio. The modified algorithm, known as the *first fit decreasing* algorithm is identical to the first fit algorithm except that items are selected for insertion in the bins in decreasing order of size. The analysis of this algorithm, which is quite long and complicated, shows that this modification results in an asymptotic performance ratio of 11/9 [5].

3. APPROXIMATION SCHEMES. Do approximation algorithms exist for all NP-complete problems? Unfortunately, it appears that the answer to this question is probably “no”. For many NP-complete problems, including the infamous traveling salesperson problem, it can be shown that the existence of a polynomial time approximation algorithm with any fixed performance ratio would imply that $P = NP$, that is, all NP-complete problems could be solved exactly in polynomial time.

On the other hand, for some NP-complete problems we can do even better than finding approximation algorithms with fixed performance ratios. For many important problems there exist families of approximation algorithms that allow us to obtain performance ratios arbitrarily close to 1 in exchange for increasingly larger polynomial time bounds. A *polynomial time approximation scheme* (PTAS) is a family of approximation algorithms $\{A_\epsilon | \epsilon > 0\}$ where for each $\epsilon > 0$, A_ϵ is a polynomial time approximation algorithm with absolute ratio bound R_{A_ϵ} at most $1 + \epsilon$.

Although it is unlikely that PTAS can be found for all NP-complete problems (since this would imply approximation algorithms for all NP-complete problems and thus that $P = NP$), it is natural to ask whether they at least exist for all problems with approximation algorithms. In a result hailed by many theoretical computer scientists as one of the most important in the field in over two decades, a group of researchers from Berkeley, Stanford, and Bell Labs showed in 1992 [1] that this too would imply that $P = NP$. Specifically, it was shown that if a PTAS exists for any problem in a rich subset of the NP-complete problems known as the MAXSNP-complete problems, then $P = NP$. Among the many important problems known to be MAXSNP-complete is the vertex cover problem.

4. FURTHER READING. Garey and Johnson’s [4] classic text offers an eminently readable introduction to NP-completeness, including a discussion of approximation algorithms and schemes. Texts by Papadimitriou and Steiglitz [8] and Cormen, Leiserson, and Rivest [3] have very good discussions and a number of illustrative examples. Motwani’s technical report on approximation algorithm [7] is also excellent. Finally, the recent result on the intractability on the hardness of

MAXSNP-complete appeared in [1] accompanied by an entertaining story in the New York Times [6].

REFERENCES

1. S. Arora, C. Lund, R. Motwani, M. Sudan, and M. Szegedy. Proof verification and hardness of approximation problems. In *Proceedings of the 33rd Annual Symposium on Foundations of Computer Science*, pages 14–23, 1992.
2. R. Bar-Yehuda and S. Even. A local-ratio theorem for approximating the weighted vertex cover problem. *Annals of Discrete Mathematics*, 25:27–45, 1985.
3. T. Cormen, C. Leiserson, and R. Rivest. *Introduction to Algorithms*. McGraw-Hill and MIT Press, 1990.
4. M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman and Company, 1979.
5. D. Johnson. *Near-Optimal Bin Packing Algorithms*. PhD thesis, Dept. of Mathematics, Massachusetts Institute of Technology, 1973.
6. G. Kolata. New short cut found for long math proofs. *New York Times*, April 7, 1992.
7. R. Motwani. Lecture notes on approximation algorithms. Technical report, Dept. of Computer Science, Stanford University, 1992.
8. C. Papadimitriou and K. Steiglitz. *Combinatorial Optimization: Algorithms and Complexity*. Prentice-Hall, 1982.

Department of Computer Science
Harvey Mudd College
301 E. 12th Street
Claremont, CA 91711
hadas@cs.hmc.edu

Mathematics, while giving no quick remuneration, like the art of stenography or the craft of bricklaying, does furnish the power of deliberate thought and accurate statement, and to speak the truth is one of the most social qualities a person can possess. Gossip, flattery, slander, deceit, all spring from a slovenly mind that has not been trained in the power of truthful statement, which is one of the highest utilities.

—S. T. Dutton

THE EVOLUTION OF . . .

Edited by Abe Shenitzer

Mathematics, York University, North York, Ontario M3J 1P3, Canada

Four Significant Axiomatic Systems and Some of the Issues Associated with Them

Stefan Mykytiuk and Abe Shenitzer

(a) Greek axiomatics and Euclid's geometry. One of the greatest intellectual achievements of the Greeks was the axiomatic method, a method for the systematic discovery of presumably absolute truths based on the application of logic to postulates and axioms. Postulates, to the Greeks, were "request(s) that something be allowed." More specifically, they were elementary, presumably obvious, truths relating to a particular discourse. (For example, the first of Euclid's postulates is: "A straight line can be drawn from any point to any point.") Axioms, to the Greeks, were elementary, presumably obvious, truths of a general nature. (For example, the first of Euclid's axioms is: "Things which are equal to the same thing are also equal to one another.") Euclid preceded his postulates and axioms with "Initial explanations and definitions" that suggest meanings and images the reader should attach to the terms of the discourse. They make it clear that one is dealing with abstractions from various physical objects. Euclid's geometry is the first known *extensive* example of what we now call an axiomatic structure.

Given some of the uses they made of Euclidean geometry, it is safe to say that the Greeks regarded it as a blueprint for the metric relations in the real world. Its uniqueness was unquestioned.

(b) Hyperbolic geometry and some effects of its discovery. Euclid relied on five postulate. The fifth of these, a kind of fly in the ointment, is the famous Euclidean parallel postulates:

If a straight line falling on two straight lines makes the interior angles on the same side together less than two right angles, the two straight lines, if produced indefinitely, meet on that side on which the angles are together less than two right angles.

"Now whatever else this postulate may be, self-evident it is not, and this was early perceived." [2] The commentator Proclus (5th century AD) objected to it by citing the asymptotic behavior of certain lines (= curves) and asking: "May not the same thing be possible in the case of straight lines . . . ?" Attempts to deduce the parallel postulate from the remaining postulates and the axioms were undertaken by various mathematicians for over 2000 years. All of them ended in failure.

A variant of Euclid's parallel postulate is the assertion: If l is a straight line and P a point not on l , then the number of straight lines through P that do not intersect l is just one. Its negation, obtained by replacing "is just one" by "is greater than one," is the so-called hyperbolic parallel postulate.

Around 1800 a few mathematicians began to experiment with the system of postulates and axioms obtained from Euclid's system by replacing the Euclidean parallel postulate by the hyperbolic parallel postulate. The latter is not an abstraction from sense impressions but a *logical* alternative to the Euclidean parallel postulate. It was Gauss, Lobachevski and Bolyai who explored the new "not-Euclidean" geometry, based on this modified foundation, in greatest depth. Lobachevski's investigations were more varied and extensive than those of either Gauss or Bolyai. Lobachevski and Gauss* carried out inconclusive physical experiments to determine which of the two geometries fitted physical space best, and formulated views of geometry radically different from the traditional ones inherited from the Greeks. Gauss' view is made clear by the following well-known quotations from his letters:

I come more and more to the view that the necessity of our geometry cannot be proved Perhaps we shall come to another insight in another life into the nature of space, which is unattainable for us now. But until then one must not rank Geometry with Arithmetic which is truly a priori, but with Mechanics . . . (From a letter to Olbers in 1817.)

It is my deepest conviction that the positions of the science of space and of the pure science of magnitude vis-à-vis our knowledge a priori differ greatly; our knowledge of the former has none of the complete conviction of necessity (and thus also of absolute truth) associated with the latter; we must admit in all humility that while number is the product of our mind alone, space has also a reality outside our mind whose laws we cannot completely prescribe a priori. (From a letter to Bessel in 1830.)

Lobachevski's sophisticated and far less well-known view is made clear by the following quotation and comment:

"In theory, nothing prevents us from assuming that the angle sum of a rectilinear triangle is less than two right angles The assumption that the angle sum of a triangle is less than two right angles is admissible only in Analytics, for measurement in nature does not reveal the slightest deflection of this sum from a half circle."

This means . . . that Lobachevski views the generalized geometry as a mental, imaginary construction which makes sense only as an analytic generalization. What justifies it is not its possible use for purposes of measurement but its usefulness for all mathematics. . . . [For him its] acceptability . . . derives from his view of a mathematical theory as a method. [5]

The first published accounts of hyperbolic geometry—by Lobachevski in 1829 and by Bolyai in 1832—had no immediate effect on the work of other mathematicians. Some of the mathematicians who were aware of the new geometric system were inclined to regard it as an aberration rather than as, in some sense, a valid alternative to Euclidean geometry. This began to change around 1860, as a result of the publication of the correspondence between Gauss and Schumacher.

*In his *Gauss, a Biographical Study* (Springer, 1981), W. K. Bühler doubts the claim that Gauss tried to determine the angular defect of a triangle determined by three mountain peaks (see p. 100).

One of Gauss' letters referred to his abiding interest in, and contributions to, hyperbolic geometry and to Lobachevski's masterly development of that geometry. Coming from Gauss, this letter generated a wave of interest in hyperbolic geometry. This interest was a key factor that ushered in a series of momentous discoveries and ideological changes not only in geometry but in all of mathematics. All this occurred in the short period between 1868 and 1872.

In 1868 Beltrami, who had familiarized himself with the work of Lobachevski, used the methods of differential geometry to establish the then surprising result that the intrinsic geometry of the pseudosphere, a surface with constant negative curvature, is *locally* hyperbolic. While doing this he also introduced an incomplete model of the hyperbolic plane in the interior of the unit disk. The gaps in Beltrami's model were filled in 1871 by Felix Klein, who arrived at *his* disk model along a projective route. *The Beltrami-Klein disk model showed that hyperbolic geometry is as consistent as Euclidean geometry. As a result, the status of Euclidean geometry as a unique system of absolute geometric truths was destroyed once and for all.*

The multiplicity of systems called geometries—Euclidean, projective and hyperbolic geometries, the geometries of surfaces in space, the geometries introduced by Riemann—gave rise to the question of what is a geometry. In his Erlangen Program of 1872 Klein gave a comprehensive (but not all-embracing) answer to this question by defining a geometry as the totality of invariants of the subsets of a set with respect to a group of permutations of that set.

“Without knowing of Klein's work, Henri Poincaré expressed similar ideas in 1880. He too was interested in [hyperbolic] geometry, and was aware of its usefulness in connection with the theory of differential equations. He picked up Beltrami's idea that on [a surface of constant curvature] it is possible to move figures without deforming them, and added that these motions form a group. Partly because of its visionary imprecision, his paper had a tremendous impact; it made the role of groups in geometry known far and wide.” [6] Some twenty years later, the demonstrated importance of groups in Galois theory, in geometry, and in analysis paved the way for group theory as a distinct area of mathematics.

Hyperbolic motions play a vital role in the theory of automorphic functions, initially developed in the 1880s by Klein and Poincaré (see Chapter 1 in [7] and the paper [8] (which also discusses recent work involving so-called hyperbolic manifolds)). They also play a key role in 4-dimensional Minkowskian geometry, the mathematical setting of the special theory of relativity first presented by Minkowski in lectures in 1905. Specifically, the group of motions of H^3 (= hyperbolic 3-space) is isomorphic to the group of homogeneous motions (= homogeneous Lorentz transformations) of Minkowskian 4-space (see Chapter 7 in [9]).

As a postscript to this account of the effects of the discovery of hyperbolic geometry we might add that it had a liberating effect not only on mathematics but also on mathematicians. As H. Weyl put it, “the individual mathematician feels free to *définir* his notions and to set up his axioms as he pleases.”

(c) Peano's axioms and “the greatest intellectual discovery of the 20th century.”

In the late 19th century, mathematicians managed to axiomatize arithmetic, and therefore, in a sense, all of mathematics. The first such axiomatization was achieved by Dedekind in 1888. Peano, working independently, published his clearer axiomatization of arithmetic a year later. This was the last triumphant step in a kind of “backward development”—from the complex to the simple. Specifically, in the 1830s Hamilton gave a rigorous definition of the complex numbers in

Descriptions of Some Technical Terms

¹**Cuts.** Split the rational numbers into two nonempty classes A and B such that every element of A is less than every element of B and such that B has no least element. Every such pair (A, B) is called a (Dedekind) *cut*.

There are natural definitions of addition and multiplication of cuts that make them into an isomorphic replica of the real number system.

²**Cardinals.** The *cardinal number* $|A|$ of a set A is, in some sense, a measure of its size. In fact, in the case of a finite set A , $|A|$ is just the number of its elements.

If there is a 1-1 correspondence between sets A and B , we write $|A| = |B|$. If A is finite or $|A| = |N|$ (where N denotes the natural numbers), then we say that A is *countable*. Otherwise A is *uncountable*.

³ and ⁵ **Well-ordered sets and ordinals.** For some ordered infinite sets, the natural numbers suffice to describe the positions of the elements. For example, in the usual ordering $1, 2, 3, \dots$ of the natural numbers, each number is both an element of the ordered set and a description of its position in the ordering.

Now consider the ordering $1, 3, 5, \dots, 2, 4, 6, \dots$ of the natural numbers. Here we run out of natural numbers after describing the positions of the odd numbers (1 is in the first position, 3 in the second, and so on). Cantor proposed the symbol ω for the position of $2\omega + 1$ for the position of 4, and so on.

Other orderings of the natural numbers led Cantor to introduce still more order symbols. For the ordering $3, 6, 9, \dots, 1, 4, 7, \dots, 2, 5, 8, \dots$ (that is, first all numbers of the form $3k$, then of the form $3j + 1$, and finally of the form $3i + 2$) he used $\omega \cdot 2$ for the position of 2 , $\omega \cdot 2 + 1$ for the position of 5, and so on.

Each of the orderings of the natural numbers which Cantor considered has the property that every nonempty subset has a least element. (Note that the integers with their usual ordering do not have this property.) He called ordered sets with this property *well-ordered*, and the new symbols he introduced to describe position in such orderings of the natural numbers, *countable ordinals*.

⁴**The continuum** is the set of all real numbers.

⁶**The axiom of choice** states that given a family of nonempty disjoint sets, a set can be constructed containing exactly one element from each set in the family.

terms of the reals, and in the 1870s Dedekind defined the reals as cuts¹ in the system of rationals, the field of quotients of the integers. Modern textbooks usually reverse the historical process and go by rigorous steps from the realm of the discrete to the realm of the continuous, from the natural numbers to the real and complex numbers.

A remarkable insight into the nature of the system of Peano's axioms, and therefore of mathematics, was achieved by Kurt Gödel in 1931. To describe it, we begin with certain preliminaries about systems of axioms.

We want a system of axioms to be *consistent*, that is, free of contradictions. If an axiom is implied by the other axioms then we can dispense with it, so it is natural to require each axiom to be *independent* of the others. Another property of a system of axioms is its *completeness*. This means that we have enough axioms to decide the truth or falseness of each statement of the system.

While we would like to know that we are working with a consistent system of axioms, we don't always strive for completeness; for example, the usual group axioms are not complete. On the other hand, it would be nice to know that Peano's axioms, the usual axiomatic basis of arithmetic, form a complete axiom set. This brings us to what is arguably the greatest intellectual discovery of the 20th

century, namely the Gödel Incompleteness Theorems. They were discovered in 1931 by the 25-year-old Kurt Gödel who proved that

For any consistent and finitely axiomatizable formal system F which contains the natural number system [with $+$ and \cdot] there are undecidable propositions in F . *One such undecidable proposition is the consistency of F .*

F. de Sua described these remarkable insights in the following witty manner:

Suppose we loosely define a *religion* as any discipline whose foundations rest on an element of faith, irrespective of any element of reason which may be present. Quantum mechanics for example would be religion under this definition. But mathematics would hold the unique position of being the only branch of theology possessing a rigorous demonstration of the fact that it should be so classified. [3]

The one island of presumed certainty of human thought was *proved* uncertain.

(d) The Zermelo-Fraenkel axiomatization of set theory and Paul Cohen's independence results. What we refer to as naive (= pre-axiomatic) set theory was greatly advanced by Georg Cantor between 1872 and 1897. Unlike the post-Zeno Greeks, Cantor accepted the actual infinite without hesitation. He made sets the ultimate components of all things mathematical and provided a calculus of sets of arbitrary "size." By the end of the century his results enjoyed wide acceptance. Then came the difficulties in the form of paradoxes (Burali-Forti, Russell, and others) and such seemingly intractable problems of set theory as the problem of the continuum hypothesis (is the cardinality² of the set of countable ordinals³ equal to the cardinality of the continuum?⁴) and the problem of well-ordering⁵ the continuum. These two problems troubled most mathematicians more than the paradoxes, which they viewed as somewhat esoteric difficulties.

The question of well-ordering the continuum was solved in 1904 by Zermelo, who showed that if one accepts the axiom of choice,⁶ then *all* sets can be well-ordered. But the axiom of choice had "side effects"—it led to various paradoxical subdivisions of figures (for example, a ball can be subdivided into five pieces (one of which is a single point) which can be reassembled into two balls each congruent to the original ball). The problem of the continuum hypothesis remained intractable.

Many of the logical difficulties associated with Cantor's set theory were overcome as a result of Zermelo's axiomatization, introduced by him in 1908 and later refined by A. Fraenkel, T. Skolem, and Zermelo himself. While its consistency is unprovable (because it effectively includes Peano's axioms), it is accepted by most mathematicians as a foundation for all mathematics more basic than Peano's axioms. The axiom of choice is now generally accepted. The continuum hypothesis remains open. In 1938, Gödel showed that these two are consistent both with each other and with the other axioms of set theory. In 1963 Paul Cohen did the same for their independence. This means, among other things, that mathematicians are free to adopt different mathematics! [4]

(e) A summary. It is useful to juxtapose key past and present views.

Until the discovery of hyperbolic geometry it was thought that postulates and axioms are abstractions from experience and, together with their logical consequences, are at least approximately true of certain objects in the real world. Consistency of the postulates and axioms was taken for granted. These were "gut feelings" as well as "official" views.

The discovery of hyperbolic geometry initiated revolutionary changes in these views of a factual as well as of a philosophical nature. There are now many axiomatic systems, including a whole hierarchy of set theories. The “official” view of postulates (or axioms—we now use the terms interchangeably) is that they are assumptions about some undefined primitive terms, hence results based on them are relative *logical* truths devoid of any outer physical meaning. The consistency of mathematics, whether we base it on Peano’s axioms or on the Zermelo-Fraenkel axioms, is *in principle* unprovable. Just as the discovery of the independence of the parallel postulate split geometry in two, so too, more than a century later, the discovery of the independence of the axiom of choice and the continuum hypothesis from one another and from the remaining axioms of set theory split mathematics. So much for facts and “official” views. Now we come to feelings.

It is safe to say that almost every mathematician is at least a “residual Platonist,” and this makes him more or less the intellectual brother of the ancient Greek mathematicians and an “emotional” opponent of formalism. Dieudonné described one variant of this syndrome in the following words:

On foundations we believe in the reality of mathematics, but of course, when philosophers attack us with their paradoxes we rush to hide behind formalism and say “mathematics is just a combination of meaningless symbols” Finally we are left in peace to go back to our mathematics and do it as we have always done, with the feeling each mathematician has that he is working with something real. This sensation is probably an illusion, but it is very convenient. That is Bourbaki’s attitude toward foundations. (Quoted in [1].)

(The Platonism of other prominent mathematicians is more robust than that of Bourbaki.)

REFERENCES

1. M. J. Greenberg, *Euclidean and non-Euclidean geometries*, Freeman, 1974 (second edition) (Chapter 8).
2. J. L. Coolidge, *A history of geometrical methods*, Dover, 1963 (Chapter IV).
3. H. Eves, *Great moments in mathematics*, the MAA, Dolciani Mathematical Expositions, volumes 5 and 7 (Lectures 7, 26, 27, 35 and 38).
4. N. Ya. Vilenkin, *In search of infinity*. (Chapter 4) A translation of this Russian book will appear in the near future.
5. V. Ya. Perminov, *The philosophical and methodological thought of N. I. Lobachevski*. (Russian, 1993. An English translation of this paper will appear in “*Philosophia Mathematica*.”)
6. *Geschichte der Algebra*, ed. E. Scholz, BI-Wissenschaftsverlag, 1990 (Section 11.4 (pp. 307-309)).
7. J. Lehner, *Discontinuous groups and automorphic functions*, AMS, 1964 (Chapter 1).
8. J. Milnor, *Hyperbolic geometry: the first 150 years*, BAMS, v.6, #1, Jan. 1982.
9. N. V. Efimov, *Higher geometry*, Mir, 1980 (Chapter 7).
10. R. Rucker, *Infinity and the mind*, Birkhäuser, 1982.

PROBLEMS AND SOLUTIONS

Edited by:
Richard T. Bumby, Fred Kochman and Douglas B. West

Proposed problems should be sent to the MONTHLY PROBLEMS address given on the inside front cover. Please include solutions, relevant references, etc. Three copies are requested.

Solutions of published problems should arrive before June 30, 1995 at the MONTHLY PROBLEMS address given on the inside front cover. Solutions should be typed with double spacing, including the problem number and the solver's name and mailing address. Two copies suffice. A self-addressed postcard or label should be included if an acknowledgement is desired.

An asterisk () after the number of a problem, or part of a problem, indicates that no solution is currently available. Partial solutions will be useful in such cases. Otherwise, the published solution is likely to be based on a solution which is complete and correct. Of course, an elegant partial solution or a method leading to a more general result is always useful and welcome. In addition, references to other appearances of MONTHLY problems or to solutions of these problems in the literature are also solicited.*

PROBLEMS

10424. *Proposed by Ira Gessel, Brandeis University, Waltham, MA.*

Evaluate the sum

$$\sum_{0 \leq k \leq n/3} 2^k \frac{n}{n-k} \binom{n-k}{2k}.$$

10425. *Proposed by Allen Barnes, Queensborough Community College, Bayside, NY.*

A circle of radius r is centered at the point $(c, 0)$. Whether or not the sine wave $y = A \sin(wx + b)$ hits the circle (i. e., touches or passes through it) depends on the values of r , c , A , w and b . Suppose that A is much larger than r and that b is chosen uniformly at random between 0 and 2π . Find the asymptotic behavior as $r \rightarrow 0$ of the probability of a hit.

10426. *Proposed by Noam Elkies, Harvard University, Cambridge, MA, and Irving Kaplansky, Mathematical Sciences Research Institute, Berkeley, CA.*

Show that any integer can be expressed as a sum of two squares and a cube. Note that the integer being represented and the cube are both allowed to be negative.

10427. Proposed by George Soules, CCR/IDA, Princeton, NJ.

Let A be an n by n positive semi-definite Hermitian matrix. Write $A = L + D + L^*$ where L is lower triangular with zero diagonal, and D is the diagonal of A (and L^* is the complex conjugate transpose of L). If $\det(D) \neq 0$, show that all n roots of $\det(zL + zD + L^*) = 0$ lie in the unit disk $|z| \leq 1$. Also, determine when this polynomial can have a root with $|z| = 1$.

10428. Proposed by Jet Wimp, Drexel University, Philadelphia, PA.

Let a_n, ϕ_n be positive constants with

$$\sum_{n=1}^{\infty} a_n \text{ convergent, and } \phi_n = O\left(\frac{1}{\log n}\right).$$

Show that $\sum_n a_n^{1-\phi_n}$ converges.

10429. Proposed by Erwin Just, Bronx Community College (Emeritus), Bronx, NY.

Let $p \equiv 1 \pmod{4}$ be a prime. Set

$$\sum_{k=1}^{p-1} (-1)^{k-1} \frac{1}{k} = \frac{A}{B} \quad \text{and} \quad \sum_{k=1}^{\frac{p-1}{4}} \frac{1}{k} = \frac{C}{D}$$

with A, B, C and D integers, and $\gcd(A, B) = \gcd(C, D) = 1$.

- (a) Prove that $p|A$ if and only if $p|C$.
- (b) Obtain an analogous result for $p \equiv 3 \pmod{4}$.
- (c) Find examples to show that these results are not *vacuously* true.

10430. Proposed by Fred Galvin, University of Kansas, Lawrence, KS, and John Isbell, SUNY, Buffalo, NY.

Let $D(a_1, \dots, a_k)$ denote the sum of the absolute deviations of the real numbers a_1, \dots, a_k from their median. Call a sequence *balanced* if the $n - 1$ quantities

$$D(a_1, \dots, a_k) + D(a_{k+1}, \dots, a_n) \quad (0 < k < n)$$

are all equal.

- (a) Show that, for each integer $n > 1$, a nonconstant balanced sequence of n terms exists, and is unique up to an affine transformation.
- (b) Characterize the positive integers n for which there exists a *strictly increasing* balanced sequence of n terms.

NOTES

(10428) Recall that $\phi_n = O(f_n)$ means that there is a constant K such that $|\phi_n| \leq K |f_n|$ for all but finitely many n (thus allowing a sloppy definition of f_n). The limits of summation, from 1 to ∞ , have not been written in the second sum, but this should cause no confusion.

(10430) For given a_1, \dots, a_k , the quantity $\sum_{i=1}^k |a_i - x|$ is minimized when x is the median of the a_i . Some examples of balanced sequences are: (length 5) 0, 2, 3, 4, 6; (length 6) 0, 1, 2, 2, 3, 4; (length 15) 0, 8, 12, 12, 14, 14, 14, 15, 16, 16, 16, 18, 18, 22, 30.

SOLUTIONS

Special Perfect Numbers

10230 [1992, 570]. *Proposed by Peter L. Montgomery, University of California, Los Angeles, CA, and John L. Selfridge, Northern Illinois University, DeKalb, IL.*

Find all perfect numbers of the form $n^n + 1$, where n is a positive integer.

Solution by Douglas Iannucci & Graeme L. Cohen, Temple University, Philadelphia, PA. The only solution is $28 = 3^3 + 1$. Let $N = n^n + 1$.

Suppose first that n is odd, so N is even. Euler proved that even perfect numbers have the form $N = 2^{a-1}(2^a - 1)$, where $2^a - 1$ is prime. Since

$$n^n + 1 = (n + 1)(n^{n-1} - n^{n-2} + \dots - n + 1),$$

and these factors are relatively prime, the even number $n + 1$ must be the factor 2^{a-1} , and therefore $2^a - 1 = 2(n + 1) - 1$. This implies $n^n + 1 = (n + 1)(2n + 1) = 2n^2 + 3n + 1$, the only solution of which is $n = 3$.

Next, suppose that n is even, so N is odd. Then n^n is a square and $n^n \equiv -1 \pmod{N}$. Thus $p \equiv 1 \pmod{4}$ for any $p|N$. In particular, N is not divisible by 3. Thus, by J. Touchard, "On prime numbers and perfect numbers", *Scripta Math.* 19 (1953), 35–39, $N \equiv 1 \pmod{12}$ and $6|n$.

Write $N = x^6 + 1$, where $x = n^{n/6} > 1$. This factors as $N = (x^2 + 1)(x^4 - x^2 + 1)$. Since $x^4 - x^2 + 1 \equiv 3 \pmod{x^2 + 1}$ and 3 divides x , these factors are relatively prime. As usual, let $\sigma(m)$ denote the sum of the divisors of m . The perfection of N implies $\sigma(N) = 2N$. Since $\sigma(m)$ is a multiplicative function, we have

$$2N = \sigma(N) = \sigma(x^2 + 1)\sigma(x^4 - x^2 + 1).$$

Since N is odd, one of the factors on the right must be odd. This implies that either $x^2 + 1$ or $x^4 - x^2 + 1$ is a square. However, $x^2 < x^2 + 1 < (x + 1)^2$ and $(x^2 - 1)^2 < x^4 - x^2 + 1 < (x^2)^2$, so there are no solutions in this case.

Editorial comment. This result for even N is explicitly mentioned in A. Mąkowski, "Remark on perfect numbers", *Elem. Math.* 17 (1962), 109.

The solution of John P. Robertson was similar to the selected solution, while Carl Pomerance and Anatoly Izotov took a different approach to the case of even n , writing $n = rs$ with $r = 2^i$ and s odd. The case $s = 1$ is ruled out by a separate argument. Otherwise, $(n^r + 1)$ is a proper divisor of N , relatively prime to its complementary factor. As above, this leads to the equation

$$\frac{x^s + 1}{x + 1} = y^2.$$

The solutions to this equation are known (see W. Ljunggren, "Noen setninger om ubestemte likninger av formen $\frac{x^n - 1}{x - 1} = y^q$ ", *Norsk. Mat. Tidsskrift* 25 (1943), 17–20) and none satisfy the other conditions of this problem.

At the 1990 Western Number Theory Conference, John Selfridge raised this question for odd N , having already solved it for even N . During the conference, Peter Montgomery solved the problem. This led to both being listed as proposers. As we have seen, the key step is to show that $3|n$ when n is even. Peter Montgomery's proof was based on the observation that $N \equiv 2 \pmod{3}$ if n is even and $3 \nmid n$, and hence, for every $d|N$, one has $d + N/d \equiv 0 \pmod{3}$. This leads to $3|\sigma(N)$, contradicting the assumptions.

Solved also by A. Izotov (Russia), C. Pomerance, and J. Robertson. Seven incomplete or incorrect solutions were received.

Applications of a Convergence Test for Fourier Series

10236 [1992, 571]. Proposed by M. J. Pelling, University College, London, England.

(a) Let $f \in L^1(\mathbb{R})$ have period 2π . Suppose that, for a given x and s , the function $\phi(u) = f(x+u) + f(x-u)$ is differentiable in an interval $(0, \delta)$, and that $\lim_{u \rightarrow 0} \phi(u) = 2s$ and $\lim_{u \rightarrow 0} u\phi'(u) = 0$. Prove that the Fourier series for f converges to s at x .

(b) Give an example for which the test in (a) succeeds while de la Vallée Poussin's test (and *a fortiori* Jordan's and Dini's tests) fails.

(c) Let $f(x) = \sum c_n x^n$ be a real power series such that $\sum c_n$ converges. By Abel's theorem, it follows that f is continuous on $[0, 1]$. Construct an example where $f(x)$ fails to be of bounded variation on $[0, 1]$.

Solution by the proposer. The three parts will be dealt with in order.

(a) In proving convergence, there is no loss of generality in assuming $s = 0$, since the general case reduces to this on replacing $f(x)$ by $f(x) - s$.

Replace δ by a smaller value if necessary, so that $\phi(u)$ and $H(u) = u\phi'(u)$ are bounded. In particular, let $|H(u)| \leq K$ and $|\phi(u)| \leq K$ for $0 < u \leq \delta$. Given $\epsilon > 0$ choose $0 < a < \delta$ so that $|H(u)|, |\phi(u)| < \epsilon$ in $(0, a]$. Consider the Dirichlet integral, $F(p) =$

$$\begin{aligned} \int_0^\delta \frac{\sin pu}{u} \phi(u) du &= \left[\frac{1 - \cos pu}{pu} \phi(u) \right]_0^\delta + \frac{2}{p} \int_0^\delta \frac{\sin^2 \frac{1}{2} pu}{u^2} \phi(u) du \\ &\quad - \frac{2}{p} \int_0^\delta \frac{\sin^2 \frac{1}{2} pu}{u} \phi'(u) du \\ &= \frac{1 - \cos p\delta}{p\delta} \cdot \phi(\delta) + \int_0^{pa/2} \frac{\sin^2 v}{v^2} \left(\phi\left(\frac{2v}{p}\right) - H\left(\frac{2v}{p}\right) \right) dv \\ &\quad + \int_{pa/2}^{p\delta/2} \frac{\sin^2 v}{v^2} \left(\phi\left(\frac{2v}{p}\right) - H\left(\frac{2v}{p}\right) \right) dv \end{aligned}$$

on substituting $u = 2v/p$. Since $\int_0^\infty (\sin^2 v) / v^2 dv = \pi/2$ it follows that

$$F(p) \leq p^{-1} \left(\frac{1 - \cos p\delta}{\delta} \right) |\phi(\delta)| + \epsilon p + 2K \int_{pa/2}^{p\delta/2} \frac{\sin^2 v}{v^2} dv \leq 4\epsilon \quad \text{for } p \geq p_0(\epsilon).$$

So $\lim_{p \rightarrow \infty} F(p) = 0$ and the Fourier series for $f(x)$ converges to s at x .

(b) If $\phi(u) = \frac{\sin \log u}{\log u}$ in $(0, \delta]$ the conditions of the test are met but we show below that $\psi(t) = \frac{1}{t} \int_0^t \phi(u) du$ is not BV (of bounded variation) in any interval $(0, a]$, hence de la Vallée Poussin's test fails.

Setting $t = e^{-v}$, $u = e^{-v-y}$, $0 \leq y < \infty$, $\psi(t) = \psi_1(v) = \int_0^\infty \frac{\sin(v+y)}{(v+y)} e^{-y} dy$ and it must be shown that $\psi_1(v)$ is not BV in any interval $[A, \infty)$. But if $v = 2k\pi$, $k \geq 1$, it is easy to see that $\psi_1(2k\pi) > B_1/k$ for an absolute constant $B_1 > 0$, and similarly $\psi_1((2k+1)\pi) < -B_2/k$ for a constant $B_2 > 0$. So the total variation of $\psi_1(v)$ in any interval $[2k\pi, \infty)$ is not less than $\sum_{n=k}^\infty (B_1 + B_2)/n = +\infty$ and hence $\psi_1(v)$ is not BV in any interval $[A, \infty)$.

(c) We prove a theorem from which examples are easily constructed. Let $h(z)$ be analytic in $|\Im(z)| < \pi/2$, $\Re(z) \leq \log 2$ and satisfy $\lim_{z \rightarrow \infty} h(z) = \lim_{z \rightarrow \infty} h'(z) = 0$ as $z \rightarrow \infty$ in this half-strip (i.e., $\Re(z) \rightarrow -\infty$). Set $f(z) = h(\log(1-z))$, $f(1) = 0$, which is defined and continuous in the disc $|z| \leq 1$, and also analytic in $|z| \leq 1$ save for the point

$z = 1$. Hence in $|z| < 1$, $f(z)$ admits a power series representation $f(z) = \sum_{n=0}^{\infty} c_n z^n$ where, by uniform continuity of $f(z)$ on $|z| \leq 1$, $c_n = \frac{1}{2\pi} \int_0^{2\pi} f(e^{i\theta}) e^{-ni\theta} d\theta$, $n \geq 0$, and $c_{-n} = 0 = \frac{1}{2\pi} \int_0^{2\pi} f(e^{i\theta}) e^{ni\theta} d\theta$, $n \geq 1$, are the Fourier coefficients of $f(e^{i\theta})$.

Theorem. The Fourier series $\sum_{n=0}^{\infty} c_n e^{ni\theta}$ converges to $f(e^{i\theta})$ at all points of $|z| = 1$, and in particular $f(1) = \sum_{n=0}^{\infty} c_n = 0$. Furthermore,

$$f(x) = \sum_{n=0}^{\infty} c_n x^n \in \text{BV}[0, 1] \iff \int_0^1 |f'(x)| dx < \infty \iff \int_0^{\infty} |h'(-x)| dx < \infty.$$

Proof. By analyticity of f on $|z| = 1$, $z \neq 1$, it follows $\sum_{n=0}^{\infty} c_n e^{ni\theta} = f(e^{i\theta})$ for $\theta \neq 0$. Near $\theta = 0$ consider $\phi(\theta) =$

$$\begin{aligned} f(e^{i\theta}) + f(e^{-i\theta}) - 2f(1) &= h\left(\log\left(1 - e^{i\theta}\right)\right) + h\left(\log\left(1 - e^{-i\theta}\right)\right) \\ &= h\left(-\left|\log\left(2\sin\frac{\theta}{2}\right)\right| - i\left(\frac{\pi}{2} - \frac{\theta}{2}\right)\right) + h\left(-\left|\log\left(2\sin\frac{\theta}{2}\right)\right| + i\left(\frac{\pi}{2} - \frac{\theta}{2}\right)\right), \end{aligned}$$

for $\theta > 0$. Apply the test of part (a) to $\phi(\theta)$. Clearly $\phi(\theta) \rightarrow 0$ as $\theta \rightarrow 0$ since $h(z) \rightarrow 0$ as $z \rightarrow \infty$. Also, $\theta\phi'(\theta) = \frac{1}{2}(\theta\cot\frac{\theta}{2} + i\theta)h'(-|\log 2\sin\frac{\theta}{2}| - i(\frac{\pi}{2} - \frac{\theta}{2})) + \frac{1}{2}(\theta\cot\frac{\theta}{2} - i\theta)h'(-|\log 2\sin\frac{\theta}{2}| + i(\frac{\pi}{2} - \frac{\theta}{2})) \rightarrow 0$ as $\theta \rightarrow 0$ since $h'(z) \rightarrow 0$ as $z \rightarrow \infty$.

So the test applies and $\sum_{n=0}^{\infty} c_n e^{ni\theta} = f(e^{i\theta})$ also at $\theta = 0$.

It is well known that if $f(x)$ is differentiable in $(0, 1)$ then $f \in \text{BV}[0, 1]$ if and only if $\int_0^1 |f'(x)| dx < \infty$. Since

$$\int_0^1 |f'(x)| dx = \int_0^1 |f'(1-x)| dx = \int_0^1 \frac{|h'(\log x)|}{x} dx = \int_0^{\infty} |h'(-x)| dx$$

the theorem follows.

Solutions to part (c) follow on taking $h(z)$ as any function satisfying the above conditions, with $\int_0^{\infty} |h'(-x)| dx = \infty$, and putting $f(x) = h(\log(1-x))$. For example, $h(z) = (\sin z)/z$, $f(x) = (\sin \log(1-x)) / (\log(1-x))$ serves since

$$\int_0^{\infty} |h'(-x)| dx \geq \int_1^{\infty} |h'(-x)| dx \geq \int_1^{\infty} \left| \frac{\cos x}{x} \right| dx - \int_1^{\infty} \left| \frac{\sin x}{x^2} \right| dx = +\infty.$$

No other solutions were received.

Almost Equidistant Vertices

10269 [1992, 958]. Proposed by D. M. Bloom, Brooklyn College, CUNY, Brooklyn, NY.

Prove that there is a constant $K < 1$ with the following property. Let \mathcal{G} be a regular $(2m+1)$ -gon inscribed in the unit circle, and let any point $P \in \mathcal{G}$ be given, then there are distinct vertices V_0 and V_1 of \mathcal{G} , such that $|d(P, V_0) - d(P, V_1)| \leq K/m$.

Solution by Robin J. Chapman, University of Exeter, Exeter, U. K. For convenience put $\alpha = 2\pi/(2m+1)$. Choose Cartesian coordinates with origin at the centre of the circumscribing circle of \mathcal{G} and such that P has coordinates $(-a, 0)$ where $0 \leq a \leq 1$. Now, there exists

a vertex V_0 of \mathcal{G} having coordinates $(\cos \theta, \sin \theta)$ with $|\theta| \leq \alpha/2$. By symmetry we may assume that $\theta \geq 0$. Let V_1 be the vertex of \mathcal{G} having coordinates $(\cos(\theta - \alpha), \sin(\theta - \alpha))$. Note that $\alpha/2 \leq |\theta - \alpha| \leq \alpha$. As

$$d(P, (\cos \phi, \sin \phi))^2 = a^2 + 2a \cos \phi + 1$$

is a decreasing function of $|\phi|$ when $0 \leq |\phi| \leq \pi$ then

$$\begin{aligned} 0 \leq d(P, V_0) - d(P, V_1) &\leq d(P, (1, 0)) - d(P, (\cos \alpha, \sin \alpha)) \\ &= a + 1 - \sqrt{a^2 + 2a \cos \alpha + 1} = f(a) \end{aligned}$$

say. Now

$$f'(a) = 1 - \frac{a + \cos \alpha}{\sqrt{a^2 + 2a \cos \alpha + 1}}$$

which is positive when $a = 0$ and can only vanish if $(a + \cos \alpha)^2 = a^2 + 2a \cos \alpha + 1$. This is impossible as $\sin \alpha \neq 0$. Thus, $f'(a) > 0$ for all $a \in [0, 1]$. Hence if $m > 1$, then $f(a) \leq f(1) = 2 - \sqrt{2} + 2 \cos \alpha = 2(1 - \cos(\alpha/2))$. Now as $m \geq 2$, $\alpha/2 \leq \pi/5 < 1$ and so $1 - \cos(\alpha/2) \leq \alpha^2/8$ (first term of Taylor expansion, which is an alternating series with terms decreasing in absolute value). Hence $|d(P, V_0) - d(P, V_1)| \leq \pi^2/(2m+1)^2$ and as the function $m \mapsto m/(2m+1)^2$ is decreasing for $m \geq 1$ then if $m > 1$, $|d(P, V_0) - d(P, V_1)| \leq L/m$ where $L = 2\pi^2/25 < 1$.

It only remains to consider the case when $m = 1$. Divide the triangle \mathcal{G} into four equilateral triangles by drawing lines between the midpoints of the sides. If P lies inside the central triangle take V_0 and V_1 to be any distinct vertices of \mathcal{G} . We may assume that $d(P, V_0) \geq d(P, V_1)$. Now $d(P, V_0) \leq 3/2$ (when P is the midpoint of the side of \mathcal{G} opposite V_0), and $d(P, V_1) \geq 3/4$ (when P is the midpoint of the side of central triangle nearest V_1). Hence $|d(P, V_0) - d(P, V_1)| \leq 3/4 < 1$. Now if P lies in the small triangle containing the vertex V of \mathcal{G} let V_0 and V_1 be the other vertices of \mathcal{G} and assume that $d(P, V_0) \geq d(P, V_1)$. Now $d(P, V_0) \leq \sqrt{3}$ (when $P = V$), and $d(P, V_1) \geq \sqrt{3}/2$ (when P is the midpoint of VV_1). Hence $|d(P, V_0) - d(P, V_1)| \leq \sqrt{3}/2 < 1$. In general

$$|d(P, V_0) - d(P, V_1)| \leq K/m$$

where $K = \max(2\pi^2/25, 3/4, \sqrt{3}/2) = \sqrt{3}/2$.

Note that the analysis for $m \geq 2$ does not give the best possible result as this method works for all P inside the unit circle, and not just for P in \mathcal{G} . One can get a *slightly* better result with a *lot* more work. Note that the bound obtained is $O(m^{-2})$. Similarly, the result for the triangle could be improved by maximizing $|d(P, V_0) - d(P, V_1)|$ over a suitable set instead of working with the individual distances.

Editorial comment. As noted when the problem was proposed, this problem sought to sharpen an upper bound of the form $1/m - A/m^3$ found in Problem A-5 on the 1989 Putnam examination. In fact, the statement was far from sharp, since upper bounds of the form Km^{-2} can be obtained. Larry Crone and Richard Holzsager observed that it is not necessary to restrict to polygons with an odd number of sides, and showed that there are vertices V_0 and V_1 of a regular n -gon inscribed in the unit circle such that

$$|d(P, V_0) - d(P, V_1)| < \frac{\pi^2}{n^2}$$

for all P in the unit circle.

Solved also by L. Crone & R. Holzsager, M. Golomb, O. P. Lossers (The Netherlands), GCHQ Problem Solving Group (U. K.), and the proposer.

Perimeters of Inscribed Polygons

10275 [1993, 75]. *Proposed by Murray S. Klamkin and A. Liu, University of Alberta, Edmonton, Alberta, Canada.*

Let \mathcal{A} be a regular n -gon with edge length 2. Denote the consecutive vertices A_0, \dots, A_{n-1} and introduce A_n as a synonym for A_0 . Let \mathcal{B} be a regular n -gon inscribed in \mathcal{A} with vertices B_0, \dots, B_{n-1} where B_i lies on $A_i A_{i+1}$ and $|A_i B_i| = \lambda < 1$ for $0 \leq i < n$. Also let C_i be the point on $A_i A_{i+1}$ with $|A_i C_i| = \alpha_i \leq \lambda$ for $0 \leq i < n$ and let \mathcal{C} denote the n -gon, also inscribed in \mathcal{A} , with vertices C_0, \dots, C_{n-1} .

With $P(\mathcal{F})$ denoting the perimeter of the figure \mathcal{F} , prove that $P(\mathcal{C}) \geq P(\mathcal{B})$.

Solution I by Albert Nijenhuis, University of Pennsylvania (Emeritus), Philadelphia, PA, and University of Washington, Seattle, WA. We show below that $\text{Area}(\mathcal{C}) \geq \text{Area}(\mathcal{B})$. \mathcal{B} and \mathcal{C} are both n -gons, and \mathcal{B} has minimal perimeter for its area, while \mathcal{C} may or may not have this property. Since minimal perimeter for n -gons is an increasing function of area, it follows that $P(\mathcal{C}) \geq P(\mathcal{B})$.

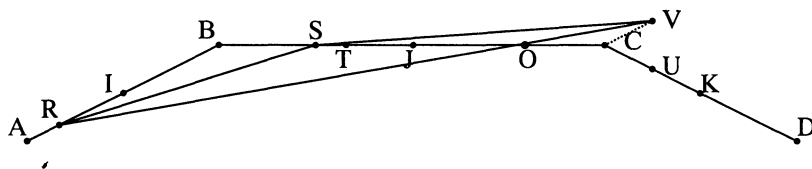
Lemma. $\text{Area}(\mathcal{C}) \geq \text{Area}(\mathcal{B})$.

Proof. We replace \mathcal{C} by a sequence of at most n n -gons, each of them obtained from the previous by replacing a vertex C_i by B_i ; the new vertex is then denoted C_i . Such a replacement is area-reducing. Indeed the half lines $\overrightarrow{A_i A_{i+1}}$ and $\overrightarrow{B_{i-1} B_{i+1}}$ intersect; as a consequence $\overrightarrow{A_i A_{i+1}}$ and $\overrightarrow{C_{i-1} C_{i+1}}$ intersect. It follows that B_i is closer to $C_{i-1} C_{i+1}$ than C_i is, so $\text{Area}(\triangle C_{i-1} C_i C_{i+1}) \geq \text{Area}(\triangle C_{i-1} B_i C_{i+1})$.

Solution II by Roy Barbara, Lebanese University, Fanar, Lebanon. First we formulate a method for comparing lengths.

Lemma. Let $ABCD$ be a convex broken line. Assume $AB = CD$ and that the angles at B and C are equal. Denote by I, J and K the midpoints of AB, BD and CD , respectively. Let R be between A and I , T between B and J , and U between C and K . Let S also lie on BT . Then $RS + SU \geq RT + TU$.

Proof.



Let V be the reflection of U across BC . Denote by O the intersection of RV and BC . Using similar triangles, it is clear that O is between J and C . Thus, T is inside the triangle RSV . Therefore: $RS + SU = RS + SV \geq RT + TV = RT + TU$.

Now we apply the lemma to solve the problem. Consider the n -gon \mathcal{C} : a first application of the lemma to $A_{n-1} A_0 A_1 A_2$ (R, S, T, U being C_{n-1}, C_0, B_0, C_1 resp.) means that replacing the vertex C_0 by B_0 will decrease the perimeter of \mathcal{C} . More generally, if we denote by \mathcal{F}_i the n -gon with vertices $B_0, \dots, B_i, C_{i+1}, \dots, C_{n-1}$ ($0 \leq i \leq n-1$), by repeated use of the lemma, we obtain $P(\mathcal{C}) \geq P(\mathcal{F}_0) \geq P(\mathcal{F}_1) \geq \dots \geq P(\mathcal{F}_{n-1})$. Since \mathcal{F}_{n-1} is \mathcal{B} , the proof is complete.

Note that we proved a more general result: the n -gon \mathcal{B} need not be regular; it is only necessary that $|A_i B_i| < \frac{1}{2}|A_i A_{i+1}|$.

Solved also by J. Fukuta (Japan), O. P. Lossers (The Netherlands), H. M. Marston, A. D. Melas (Greece), R. M. Robinson, A. A. Tarabay (Lebanon), A. Tissier (France), J. C. Vera Lizcano (student, Colombia), A. N. 't Woord (The Netherlands), and the proposer.

Collaborating editors: *David F. Appleyard, Paul T. Bateman, Bruce C. Berndt, Duane M. Broline, Barry W. Brunson, Frank S. Cater, Gulbank D. Chakerian, Underwood Dudley, Gerald A. Edgar, Michael A. Filaseta, Ira M. Gessel, Richard A. Gibbs, Jerrold R. Griggs, Douglas A. Hensley, John R. Isbell, Mourad E. H. Ismail, Murray Klamkin, Daniel J. Kleitman, Frederick W. Luttman, Frank B. Miles, Richard Pfiefer, Stephen L. Portnoy, J. O. Shallit, John Henry Steelman, Kenneth B. Stolarsky, David E. Tepper, Douglas B. Tyler, Daniel Ullman, and William E. Watkins.*

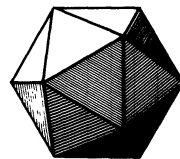
Carl Runge on Algorithmics

Suppose the mathematician gives them a method of calculation perfectly logical and conclusive but taking 200 years of incessant numerical work to complete. They would be justified in thinking that this is not much better than no method at all. So there arises a third stage of the solution of a mathematical problem in which the object is to develop methods for finding the result with as little trouble as possible. I maintain that this third stage is just as much a chapter of mathematics as the first two stages, and it will not do to leave it to the astronomer, to the physicist, to the engineer or whoever applies mathematical methods, for this reason that these men are bent on the results and therefore they will be apt to overlook the full generality of the methods they happen to hit on, while in the hands of the mathematician the methods would be developed from a higher standpoint and their bearing on other problems in other scientific inquiries would be more likely to receive the proper attention.

From *Graphical Methods* by Carl Runge,
Columbia University Press, 1912.

Submitted by Steve Maurer

The American Mathematical Monthly



Volume 102 Number 2 / FEBRUARY 1995



Anneli Lax

AN OFFICIAL PUBLICATION OF THE MATHEMATICAL ASSOCIATION OF AMERICA

NOTICE TO AUTHORS

The *Monthly* publishes articles, notes, and other features about mathematics and the profession. The readership of the *Monthly* is intended to include everybody who is mathematically inclined, including of course professional mathematicians and students of mathematics at all collegiate levels. While no single article or feature is likely to appeal to everyone, material should interest and be accessible to a large number of readers. This is the most important criterion for acceptance.

Articles may be expositions of old results or presentations of new ones. They may concern all of mathematics or one small area, a broad development or a single application, historical reminiscences or one important event. While some articles may contain the author's new research, the novelty of material and generality of the results is far less important than the clarity of exposition and general interest. Discussing one illuminating case of a well known result is far better than providing all the details of an obscure but new proposition. Articles in the *Monthly* are supposed to inform and to entertain; they are meant to be read rather than archived.

Notes are short and possibly informal articles. A note may concern a clever new proof of an old theorem, a novel way to present tired material, or a lively discussion of a philosophical (but still mathematical) issue. Also, any topic is suitable, so long as it is related to mathematics. Because a note is short, the first few sentences are the most important part: They should explain the purpose and invite the reader in. Photographs or diagrams often will attract the reader's attention.

All articles and notes should be sent to the editor:

JOHN EWING
Department of Mathematics
Indiana University
Bloomington, IN 47405

Please send 3 copies, typewritten on only one side of the paper. Illustrations should be carefully drawn on separate sheets of paper in black ink; the original should be without lettering and two copies should have appropriate captions and lettering indicated.

Proposed problems or solutions should be sent to:

RICHARD BUMBY,
P.O. Box 10971
New Brunswick, NJ 08906-0971.

Please send 2 copies of all material, typewritten if possible.

Letters to the Editor, both for publication and for private reading, should be sent to the Editor at the address given above. Comments, including criticisms, are welcome, as are all suggestions for making the *Monthly* a lively, entertaining, and informative journal.

EDITOR:

JOHN H. EWING

ASSOCIATE EDITORS:

PETER BORWEIN	FRED KOCHMAN
RICHARD BUMBY	CATHERINE MCGEOCH
DENNIS DETURCK	RICHARD NOWAKOWSKI
UNDERWOOD DUDLEY	ARNOLD OSTEBEE
JOHN DUNCAN	LEE RUBEL
JOAN FERRINI-MUNDY	ABE SHENITZER
JOSEPH GALLIAN	LYNN STEEN
STEVEN GALOVICH	STAN WAGON
RICHARD GUY	DOUGLAS WEST
DARRELL HAILE	HERBERT WILF
PAUL HALMOS	SANDY ZABELL
JOAN HUTCHINSON	PAUL ZORN

EDITORIAL ASSISTANT:

MISTY CUMMINGS

STAFF ARTIST:

MIKE CAGLE

QUOTE MASTER:

MARK WOODARD

Reprint permission:

MARCIA P. SWARD, Executive Director

Advertising Correspondence:

Ms. ELAINE PEDREIRA, Advertising Manager

Subscription correspondence, change of address, and other inquiries:

Membership / Subscriptions Department

All at the address:

The Mathematical Association of America
1529 Eighteenth Street, N.W.
Washington, DC 20036.

Microfilm Editions: University Microfilms International, Serial Bid coordinator, 300 North Zeeb Road, Ann Arbor, MI 48106.

The AMERICAN MATHEMATICAL MONTHLY (ISSN 0002-9890) is published monthly except bimonthly June-July and August-September by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, DC 20036 and Montpelier, VT. Copyrighted by the Mathematical Association of America (Incorporated), 1995, including rights to this journal issue as a whole and, except where otherwise noted, rights to each individual contribution. General permission is granted to Institutional Members of the MAA for noncommercial reproduction in limited quantities of individual articles (in whole or in part) provided a complete reference is made to the source. Second class postage paid at Washington, DC, and additional mailing offices. **Postmaster:** Send address changes to the American Mathematical Monthly, Membership / Subscription Department, MAA, 1529 Eighteenth Street, N.W., Washington, DC, 20036-1385.

**The American
Mathematical Monthly**

Volume 102 Number 2 / FEBRUARY 1995
(ISSN 0002-9890)



Contents

ARTICLES

- Yueh-Gin Gung and Dr. Charles Yu Award for Distinguished Service to
Anneli Lax / IVAN NIVEN 99
- Calculus in the Operating Room / PEARL TOY and STAN WAGON 101
- Experimentation and Conjecture Are Not Enough / DEBORAH TEPPER
HAIMO / 102
- Pebbling a Chessboard / FAN CHUNG, RON GRAHAM,
JOHN MORRISON, and ANDREW ODLYZKO 113
- Drums That Sound the Same / S. J. CHAPMAN 124
- Down With Determinants! / SHELDON AXLER 139
-

FEATURES

COMMENTS 98

NOTES

- Where Not to Find the Critical Points of a Polynomial—Variation
on a Putnam Theme / PETER ANDREWS 155
- A Short Path to the Shortest Path / PETER D. LAX 158
- A Note on Entire Solutions of the Eiconal Equation /
DMITRY KHAVINSON 159
- The Uniqueness Aspect of the Fundamental Theorem of Finite
Abelian Groups / DAVID B. SUROWSKI 162

UNSOLVED PROBLEMS

- Coin-Weighing Problems / RICHARD K. GUY and
RICHARD J. NOWAKOWSKI 164

THE AUTHORS 168

PROBLEMS AND SOLUTIONS 169

REVIEWS

- Politics, Logic, and Love: The life of Jean van Heijenoort.*
By Anita Burdman Feferman / JEFFERY NUNEMACHER 178

TELEGRAPHIC REVIEWS 180

Calculus in the Operating Room

Pearl Toy, M.D., and Stan Wagon

Here is a realistic, and potentially important, application of a familiar topic from calculus. Imagine a hospital patient about to undergo surgery. Suppose he has 5 liters (L) of blood in his body, 40% of which consists of red blood cells (this percentage is called the *hematocrit*), and, during the surgery, he will bleed $2\frac{1}{2}$ liters of blood. This is a realistic estimate for certain types of hip replacement, for example. His blood volume is maintained at 5 L by controlled injection of saline solution (no blood cells), which we assume to mix instantaneously with his blood. This means that the blood lost through bleeding becomes less rich in red cells as the operation progresses.

Question 1. What is the patient's volume of red blood cells at the end of the operation?

Some of the lost blood can be recovered, washed, and returned to the patient after the operation; but there is some loss and washing is an expensive procedure. Suppose that, before the operation, some blood is removed from the patient and replaced with saline solution. This blood will be returned to the patient afterward. This procedure, called acute normovolemic hemodilution (ANH), will decrease the loss of red blood cells during the operation. However, during the transfusion the patient's total blood volume is maintained at 5 L; as with the bleeding during surgery, this affects the rate of red blood cell removal.

Question 2. If it is known that the patient's hematocrit can go as low as 20%, but no lower, how much blood should be replaced in the ANH procedure just described?

Both of these questions can be answered by a simple exponential decay model, once one makes the observation that the rate of blood cell loss during the operation is proportional to the amount of red cells present. For the numbers given, and with time measured as a fraction of the length of the operation, the proportionality constant is $\frac{1}{2}$, since 2.5 of the 5 L are lost.

Now both questions can be answered. If $f(t)$ is the volume of red blood cells remaining at time t , then $f(t) = f(0)e^{-t/2}$. For Question 1, $f(1) = 2000/e = 1213$ milliliters.

For Question 2 we need to know what value of $f(0)$ will cause $f(0)/e$ to be 1000 ml (20% of 5 L); this is $f(0) = 1000e$. In order to figure out how much should be removed by ANH, we reverse the technique just discussed and solve $2000e^{-k/5000} = 1000/e$ to get $k = 966$ ml. This leaves $1000/e$, or 1649 ml of red blood cells in the patient for a hematocrit of 33%, which will become 20% after surgery.

Note that without the transfusion the red blood cell loss is $2000 - 1213 = 787$ ml. With the transfusion the patient starts with 1649 ml of red blood cells and ends up with 1000; a loss of 649 ml. There is a net savings of 138 ml of red blood cells. This savings may in fact not be large enough to justify the procedure; it must be balanced with overall expense, risks associated with ANH, and the risk of an adverse reaction to blood the patient may have to receive from a blood bank after the operation (see M. E. Brecher and M. Rosenfeld, *Mathematical and computer modeling of acute normovolemic hemodilution, Transfusion* 1994 (34), 176-179). But it is noteworthy that a simple freshman-calculus model applies to the basic situation.

Moffitt-Long Hospital
University of California
San Francisco, CA

Department of Mathematics
Macalester College
St. Paul, MN

Experimentation and Conjecture Are Not Enough

Deborah Tepper Haimo

Dedicated to the Memory of Franklin Tepper Haimo

This is an exciting time in mathematics. Its various special areas are coming together to emphasize the discipline's unity. In addition, there is general growing recognition that good teaching cannot be separated from research, and that we must be more successful in communicating the tenets of our field to the broader community.

In underscoring the beauty of mathematics as well as its relevance, it is of utmost importance that we educate those naturally gifted and interested, as well as others who may have latent ability which should not be ignored, and the many who need to learn at least the basic concepts and to gain some appreciation and understanding of the field's vitality.

It is heartening to see that we are beginning to acknowledge our responsibility to become involved in raising the understanding of the mathematical knowledge of all our citizenry. We need

- creative investigators at the forefront of research to advance the field through major breakthroughs,
- our professionals to apply developed theories to practical applications,
- our instructional sector at every level, from kindergarten teachers to university research professors, to educate our youth,
- our amateurs to generate and maintain public interest in our field,
- and a broad, mathematically literate citizenry to appreciate its importance and power and its need for public support.

It is daunting to contemplate how extensive a challenge we face, and how much there is to accomplish.

We may recall that, not very long ago, mathematics was described, on the one hand, as the queen of the sciences, and on the other, as the handmaiden of the sciences. It is interesting to note that, in these characterizations, our discipline is identified as "feminine", although there seems to be some question about her social status, ranging from one extreme to the other.

Today, mathematics is generally viewed on a more neutral level with the sciences, neither in a lofty regal position nor in a lowly subservient one. It is important, however, that it continue to maintain its unique quality.

As we enter an era of great change, we need to recognize and stress the distinct nature of mathematics that differentiates it from all other sciences. In particular, as we return, in our educational approach, to the historic emphasis on problem-solving, along with our current focus on experimentation and conjecture, we must

be careful to appreciate the fact that, laudable as this approach is, it is a start, but not enough!

While the introduction of problem-solving is intended to involve students more actively in their learning and understanding of the nature of mathematics,—and that is certainly a positive factor,— there is another vital aspect, the need for proof, unique to mathematics, and yet all too often, ignored in our educational process. Problem-solving is not complete until the results have been firmly established. Proofs are an integral part of mathematics and must not be overlooked!

It is important that, in our efforts to make mathematics more relevant, attractive, and accessible to a greater number of students, we avoid discarding its fascinating and surprising features. Since these features must have attracted most of us, it seems reasonable to believe that they would appeal also to a segment of our students.

We cannot afford to neglect the education of those students who are drawn to what generally are thought to be the more challenging aspects of mathematics; we cannot assume that they will manage well without our help and direction. As one of them once plaintively exclaimed “just because some of us are considered able doesn’t mean that we don’t need guidance; we flounder and become discouraged too!”

In order that we have better success in interesting most students, problem-solving with experimentation and conjecture, may be a reasonable approach. At the same time, we must also expose our students to proofs, and must strive for a reasonable balance. We must be careful not to distort the subject by removing its major distinguishing component and eliminating the essence of its uniqueness. Whereas in some areas, substantiation of a theory for a large number of instances is considered adequate for acceptance, not so in mathematics.

I remember that when I was a mathematics undergraduate many years ago, we used to consider physics majors as coming to conclusions that were invariably sloppy. Our favorite description of physicists, as distinguishable from mathematicians, was that they believe that all odd numbers are prime, and that they reach that result by the argument that 3 is a prime, 5 is a prime, 7 is a prime, 9 is an experimental error, 11 is a prime, . . . , q.e.d.

Of course we exaggerated, but we had a point that always applies to mathematics. In mathematics, there is no middle ground; it is “all or nothing”! Looking at a number of examples and recognizing relationships among elements may lead to interesting and useful conjectures, but they remain no more than that—or, to quote a familiar refrain, “tain’t necessarily so”. A conjecture becomes accepted as fact only after it has been confirmed by an acceptable proof, or else a counterexample must be found to reject it.

In the current reform movement in mathematics education, there is recognition of the importance of stressing that experimentation and conjecture lie at the very core of the field. There is also mention of the need for justification in some form, at appropriate levels, before a presumed conclusion can be accepted. Unfortunately, particularly in introductory courses, the problem-solving aspects are emphasized greatly while proofs of results are generally barely mentioned or ignored entirely.

The stress on problem-solving, where students are encouraged to look for patterns and draw conclusions, merits our applause. That is the nature of mathematics; that is when experimentation and conjecture occur. It is, however, only a beginning, and this must be made unmistakably clear if we are serious about educating our students fully.

Depending on the educational context, a rigorous proof may not be in order, nor may it be necessary initially to present more than an intuitive outline. Both teacher and student must be thoroughly aware, however, of the limitations of what has been accomplished—no more than the validity of the finite number of special cases considered.

What appears to hold and what seems totally plausible may actually not be so in general. The crux of mathematics is lost if that fact is not clearly understood and appreciated, and all our students must recognize this. If we are to educate them, we need to convey to all the essence of mathematics. It is imperative that we not only address the needs of those who seek relevance, but also, maintain the interest of those able to understand the abstract nature of mathematics and to appreciate its beauty; these students, too, must be encouraged to use their abilities to the fullest to continue as leaders to the forefront of our field.

As our students engage in problem-solving, we must point out that examples abound of conjectures that fail to hold beyond some point. Let us recall, for example, Euclid's indirect proof, as generally given in an introductory course in algebra, that there are an infinite number of primes. That proof might suggest that a way to construct primes is to multiply all those up to a certain value and then to add 1. We thus would start with the first prime 2 to get

$$2 + 1 = 3;$$

and would continue, getting the following successive results

$$2 \cdot 3 + 1 = 7;$$

$$2 \cdot 3 \cdot 5 + 1 = 31;$$

$$2 \cdot 3 \cdot 5 \cdot 7 + 1 = 211;$$

$$2 \cdot 3 \cdot 5 \cdot 7 \cdot 11 + 1 = 2,311.$$

Since 3, 7, 31, 211, and 2311 are all primes, a conjecture that this algorithm will generate only primes would seem reasonable at this point. Unfortunately, we next have

$$2 \cdot 3 \cdot 5 \cdot 7 \cdot 11 \cdot 13 + 1 = 30,031,$$

and

$$30,031 = 59 \cdot 509,$$

so that conjecture fails on the sixth attempt.

In mathematical work, experimentation often with specific concrete examples to test a conjecture, is the norm, though this is generally hidden from sight. As one faculty member once remarked, "mathematicians' experiments end in the waste basket!". Divulged are not the false starts and the difficulties encountered, nor the messy computations, but only the carefully executed final results. Such omissions, sometimes without even any public indication of the insight and motivation that led to a particular conclusion, is regrettable, and helps to create an aura of an incomprehensible subject.

Students, for the most part, engage in solving problems whose solutions are known, even if not to them. It is thus of utmost importance that teachers, especially, be in a position to provide useful guidance. They must be knowledgeable of the existence of a proof or a counterexample for a given problem; otherwise, they may not realize that their students' experiments may not have been carried out far enough to determine whether there is a valid basis for an algorithm that holds in general for that problem.

We are able to disprove our first example rather quickly, but not all conjectures can be so readily discarded. That is why it is essential to establish general validity.

Let us turn, for example, to a problem proposed by George Polya, acknowledged as one of the most renowned of problem-solvers.

In 1919, Polya made an interesting observation. He let

n represent any positive integer,

r , the number of prime factors of n , counting multiplicities, and taking $r = 0$ when $n = 1$, and $r = 1$ when n is a prime,

O_x , (x an integer), the number of positive integers $\leq x$ with an odd number of prime factors,

E_x , the number of positive integers $\leq x$ with an even number of prime factors.

Polya conjectured that, for $x \geq 2$,

$$O_x \geq E_x.$$

Further, if

$$L(x) = E_x - O_x,$$

then

$$L(x) \leq 0.$$

The function, L , can be written as a sum

$$L(x) = \sum_{n=1}^x \lambda(n), \quad x > 1,$$

where $\lambda(n)$ is the Liouville function

$$\lambda(n) = (-1)^r.$$

The Polya conjecture is not an arbitrary exercise, but an attempt to relate the Liouville function to the distribution of prime factors. Its validity for the first 50 consecutive positive integers is readily established by hand, and many, many further cases confirm this observation. Indeed, the conjecture was generally deemed true for nearly 40 years, until 1958, when C. B. Haselgrove proved that $L(x) > 0$ for infinitely many x . He failed, however, to produce any specific x for which the conjecture fails.

In 1962, R. S. Lehman found that

$$L(906, 180, 359) = 1,$$

and in 1980, M. Tanaka discovered that the smallest counterexample of the Polya conjecture occurs when

$$x = 906, 150, 257.$$

Thus a very promising conjecture fell by the wayside! Note, however, that this did not occur for the first 906, 150, 256 integers! Indeed, although a counterexample may not appear before the trillionth case, or even much later, it is enough for disqualification.

Mathematics provides for other disciplines the compactness and simplicity of language that allows useful descriptions of fundamental results and natural phenomena. Indeed, to most of the outside world, the importance of mathematics lies in its utilitarian role.

Those outside the discipline rarely appreciate the subject's intrinsic beauty nor marvel at its great power. That power, for such broad and diverse applications, is derived from its abstract nature, the very characteristic that instills awe, but alienates many of those merely interested in the concrete and the applicable.

While the need to make the subject more relevant may be inescapable, it should not be so all encompassing that the essence of mathematics as a major discipline in its own right is totally lost. It is a field with problems that attract and fascinate some with no interest in the applied or the relevant.

Many of those who have worked in such areas as number theory had no intent to solve practical problems. Yet, as we know, their results have sometimes produced incredible applications many years later. We must thus make sure that we do not fail to captivate and nurture students with such a bent who may later contribute to the advancement of the discipline itself. These will be the researchers of tomorrow who will continue to develop the theoretical basis which sometimes leads to unexpected and important applications that benefit all of us.

There are many examples in number theory of conjectures that fail to hold. These generally have a universal appeal, since they involve positive integers, and are readily understood. As is well known to mathematicians, many are extremely difficult to prove conclusively. Indeed, attempts to prove some seemingly simple problems, using known techniques, are often fruitless. In some cases, however, these concerted efforts have led to the creation of new methods and the introduction of new fields of far greater significance and impact, both to mathematics and its applications, than the original rather specialized quest might have indicated.

A famous example, of course, is Fermat's last theorem. When, after some 350 years, a proof was recently announced by Andrew Wiles of Princeton University, the imagination of the general public was aroused to such an extent that a popular presentation, arranged at the San Francisco Exploratorium, sold out quickly and even had scalpers charging many times the regular cost of a ticket!

Simply stated, the Fermat conjecture generalizes the well known Pythagorean theorem by asserting that there are no positive integers x , y , z , and $n > 2$, such that

$$x^n + y^n = z^n.$$

As is well known, Fermat claimed to have found what he described as a "marvellous" proof of his conjecture that there exist no non-trivial solutions to the problem. He explained that he omitted the proof due to his inability to fit it into the margin of his book. There are serious doubts that his proof was valid. Indeed, it would take more than Fermat's book margin just to list the names of all those who have made significant contributions in the course of trying to prove the conjecture.

In his attempts to settle the question, Wiles himself appealed to results of elliptic curves, Galois representations, and modular forms in order to make substantial headway. His announced version still is not entirely complete. He expects, however, that in the near future, he will have the needed final computation of a precise upper bound for the Selmer group in the semistable case.

Some of the techniques already developed in earlier years have proved useful in dealing with related problems. For example, after establishing, in 1769, that the special Fermat equation

$$x^3 + y^3 = z^3,$$

has no non-trivial integer solutions, Euler made a more general conjecture surmising that the equation

$$x_1^n + x_2^n + \cdots + x_{n-1}^n = x_n^n$$

has no positive integer solutions x_1, x_2, \dots, x_n .

For $n = 5$, a counterexample was produced by a direct computer search in 1966 by L. J. Lander and T. R. Parkin. They found that

$$27^5 + 84^5 + 110^5 + 133^5 = 144^5.$$

It is interesting to note that, when $n = 4$, the conjecture could neither be proved nor could a counterexample be found by a similar direct computer search. In 1987, however, Noam Elkies of Harvard University introduced a pencil of curves of genus 1 which lies on

$$a^4 + b^4 + c^4 = d^4,$$

a, b, c, d integers, and found the simplest curve in the pencil which could possibly lead to a rational point that would disprove the Euler conjecture. He then used the computer and succeeded in establishing that

$$2,682,440^4 + 15,365,639^4 + 18,796,760^4 = 20,615,673^4.$$

In disproving the Euler conjecture for $n = 4$, he not only established the result, but demonstrated the importance of mathematical analysis to enable him to take advantage of the computer. It was only after restricting the variables to lie on an appropriate curve that the solution was found, being beyond the range of earlier exhaustive searches.

With this solution revealed, Elkies was then able to exploit the theory of elliptic curves to generate recursively arbitrarily many other solutions from that one. In a subsequent computer search, a minimal solution of the equation was found by Roger Frye of Thinking Machines to be

$$95,800^4 + 217,519^4 + 414,560^4 = 422,481^4.$$

Aside from number theory, elementary problems in other areas may also lead to erroneous conclusions if experimentation is terminated prematurely and no proof ensues. As an example, consider n points on a circle, connected pairwise by chords, no three of which are concurrent within the circle. The problem is to determine the number of distinct, non-overlapping regions, R_n , formed.

We start by counting, and note that

$$R_1 = 1$$

$$R_2 = 2$$

$$R_3 = 4$$

$$R_4 = 8$$

$$R_5 = 16$$

strongly suggesting the reasonable general algorithm

$$R_n = 2^{n-1}, n \geq 1.$$

To check this result, we turn to R_6 and find

$$R_6 = 31.$$

Our careful count turned up no more than 31 regions, one short of the number predicted by our conjecture.

Fortunately, we can analyze the situation mathematically, and note that we can invoke the familiar Euler formula

$$F = 1 - V + E.$$

Here we have the number of faces of a planar map in terms of the number of

vertices and edges that determine those faces. Now,

the faces F are the R_n regions;

the vertices V are the n points on the circle and the interior points i resulting from intersections of chords;

the edges E are our n circular arcs and the chord segments formed by the interior points.

Substituting into the Euler formula, and using an identity involving the binomial coefficients, we have

$$R_n = C(n-1, 0) + C(n-1, 1) + C(n-1, 2) + C(n-1, 3) + C(n-1, 4),$$

the sum of the first 5 binomial coefficients in the expansion of

$$\begin{aligned} 2^{n-1} &= (1 + 1)^{n-1} \\ &= C(n-1, 0) + C(n-1, 1) + C(n-1, 2) + \cdots + C(n-1, n-1). \end{aligned}$$

It is now clear that the two agree only when $n \leq 5$.

Pictorially, we have the Pascal triangle, with a section deleted after the fifth row of numbers, as shown below:

$$\begin{array}{ccccccccccc} & & & & 1 & & & & & & \\ & & & 1 & & 1 & & & & & \\ & & 1 & & 2 & & 1 & & & & \\ & 1 & & 3 & & 3 & & 1 & & & \\ & 1 & 4 & & 6 & & 4 & 1 & & & \\ & 1 & 5 & 10 & & 10 & 5 & /1 & & & \\ 1 & 6 & 15 & 20 & & 15 & /6 & 1 & & & \\ 1 & 7 & 21 & 35 & & 35 & /21 & |7 & 1 & & \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \end{array}$$

A less tractable problem in this general area is the four-color map problem that has challenged mathematicians for many years. Indeed, G. D. Birkhoff once noted that every serious mathematician that he knew, had, at one time or other, tried to prove the four-color map problem. This problem claimed that, to differentiate among the countries of any map for which no contiguous regions meet in just one point, four is the minimum number of colors needed.

First proposed in 1852 by Francis Guthrie, it gained substantial importance and attracted wide attention when a proof, published in 1879, was found to have a serious gap eleven years later. Many of the famous mathematicians who devoted much time and effort to seeking a solution, created new fields in the process, as is not uncommon.

Substantial interest in the problem continued throughout the years. Indeed, during my student days, it was one of the foremost problems discussed. I remember, particularly, that when Hassler Whitney was asked one day about how long a dissertation had to be, he responded that a two-line proof of the four color problem would suffice.

Many years later, in 1976, Kenneth Appel and Wolfgang Haken produced a proof, but it not only was far longer than a mere two lines, its techniques were totally unorthodox. After reducing the problem to consideration of the characteristics of some 2000 different maps, they programmed a computer to determine the outcome. The use of a computer to solve a mathematical problem remains

unacceptable to some, and it was expected that an analytic proof would soon be found. Now, nearly 20 years later, this had not occurred.

In recent years, as our technology has developed, the introduction of computers has extended far beyond mere computer-assisted proofs to the point that there are advocates of using computers to provide probabilistic proofs as well as visual video proofs, thus changing the very nature of mathematics which has always relied on rigorous analytic proofs.

While we must prepare our students to recognize the importance of computers, we must instruct them in their appropriate use. Using computers to suggest patterns or inspire intuition replaces old tool with new, but does not change the character of proof. To go further, however, and accept as valid results that have been shown to hold for some finite number of cases, however large, because it has not been possible to establish them in general is to engage in a different area.

The high standards of rigor, introduced after informal reasoning created serious problems in the 18th century, have served us well, propelling our field forward at an impressive rate. We should not reverse that direction.

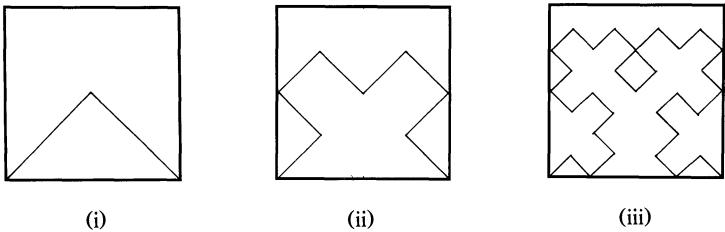
One of the intriguing characteristics of mathematics is the existence of properties that are totally counter to intuition. Some results can be convincingly proved, yet they confound the imagination as they seem to contradict what reason would seem to dictate.

Such a situation is encountered early in a calculus course. It occurs when the curve, described by $y = 1/x$, is rotated about the x -axis for $x \geq 1$. The resulting surface of revolution, known as Gabriel's horn, has the surprising property that we can fill the horn with a finite amount of paint, but no matter how many times we might refill it, we never would have enough to paint its surface, a remarkable example of a body of finite volume and infinite surface area!

Mathematics demands a high level of precision in the concepts we introduce, as their validity must be upheld when they contradict our intuitive notions. This was certainly the case in 1890 when the mathematical community was electrified by G. Peano's presentation of a space-filling curve. He established the existence of a continuous map f of the closed unit interval onto the closed unit square. One illustration is the following construction of a sequence of continuous functions f_n which converge to the limiting function f sought.

We start out by bending of the unit interval into a right isosceles triangular segment, (i) below.

Every such right isosceles triangular segment is further mapped into a chain of 4 such segments, each of $1/2$ the length, with the chain beginning at the same initial point and ending at the same final point, as in (ii). Repeating the operation on each of the four right isosceles triangular segments, we have the next member of our sequence as shown in (iii).



We continue to repeat the operation recursively noting that our path is made up of 4^n right isosceles triangular segments each lying in a square of side $1/2^n$. Each of these paths represents a continuous function f_n , and covers more of the initial square.

It is then not difficult to establish that the sequence f_n converges to a continuous surjective function f .

A computer can be used here for great effect to give a visual demonstration of the recursive procedure described. It will provide a clear picture to show the square being covered by the continuous curve progressively more and more until it appears completely covered.

Strange as space-filling curves are, an even more counter-intuitive result was established by Felix Hausdorff in 1914 when he sought to determine whether there exists a non-trivial, finitely additive, congruent-invariant measure. He succeeded in establishing the non-existence of such a measure in three dimensional space when he proved that, neglecting a denumerable set, the surface of a sphere can be decomposed into a finite number of pieces and reassembled by rigid motions to form two spheres, each with the same radius as the original one.

A decade later, Stefan Banach and Alfred Tarski proved the same property for a solid sphere, without the need to remove a denumerable set. They went even further, obtaining the more general result that, in three dimensional space, given any two bounded sets, with interior points, one of the sets can be decomposed into a finite number of disjoint subsets which can be reassembled to form the other set, with no gaps or overlaps. In short, the two sets are equidecomposable.

Banach and Tarski succeeded in showing the fascinating and nonintuitive result that, say, the earth can be decomposed into a finite number of pieces which can be reassembled to form a marble, or even to form two earths each of the same size as the original. John von Neumann added to these amazing facts the observation that only nine pieces are needed for the decomposition of one sphere into two, all with the same radii. Abraham Robinson went further yet in 1947 showing that five pieces will suffice!

In the plane, the Banach-Tarski Paradox fails to hold. We have, rather, that any two well-behaved planar sets with the same area are equidecomposable. Tarski conjectured that a closed disk can be cut up into a finite number of pieces which can be reassembled to form a square exactly, with no gaps or overlaps. This was proved in 1989 by the Hungarian mathematician Miklos Laczkovich.

Rather than the small number of pieces into which one sphere in space may be cut in order to be reassembled into two, Laczkovich found that, in the plane, to effect the decomposition of the closed disk to form a square, it was necessary to have some 10^{50} pieces of a great variety of strange shapes. Remarkably, these pieces merely had to be translated to new positions to form a square exactly.

The talent and persistence of mathematicians, using ever growing and more powerful mathematical tools, will ultimately succeed in resolving the questions posed in some of the outstanding conjectures.

Left over from earlier centuries and one of the remaining major unsolved mathematics problems, the Riemann Hypothesis, has a relationship to analysis that corresponds to that of Fermat's Last Theorem to arithmetic. If proved true, it would have significant consequences in number theory, further unifying seemingly independent areas of mathematics. The Hypothesis is particularly difficult to establish since there seems no rationale to indicate why it might be true.

For a complex variable s , Riemann defines his zeta function, ζ , by the series

$$\zeta(s) = \sum_{n=1}^{\infty} 1/n^s.$$

The Liouville function λ , which we encountered when we considered the Polya conjecture, is connected to Riemann's zeta function by the equation

$$\zeta(2s)/\zeta(s) = \sum_{n=1}^{\infty} \lambda(n)/n^s.$$

Riemann established that all zeros of $\zeta(s)$ in the right half plane lie in the unit vertical strip $0 \leq \operatorname{Re} s \leq 1$. He then stated his famous conjecture that all non-negative zeros of $\zeta(s)$ lie on the line $\operatorname{Re} s = 1/2$, an incredible conclusion for which there seems to be no discernible reason.

In 1914, G. H. Hardy proved that there are infinitely many zeros on the line $\operatorname{Re} s = 1/2$. Confirmation of the conjecture for a finite number of cases has been steadily increasing. In 1955, computers were used to substantiate the fact that Riemann's conjecture holds for the first 25,000 zeros of the zeta function, all simple and all on the line $\operatorname{Re} s = 1/2$. In 1969, it was found that this continues to be true for the first 3,500,000 zeros of the zeta function; i.e. they are all simple and lie on the line $\operatorname{Re} s = 1/2$.

These examples, finite in number as they are, provide increasingly convincing substantiation of the validity of the Riemann hypothesis. Indeed, the use of high speed computers has confirmed the conjecture for the first 1.5 billion zeros. Nonetheless, problems may arise as soon as computation lies beyond the range of any existing computer. That is why mathematics requires solid proofs and why the new technology can be so helpful, if used properly, either in providing counterexamples or in aiding in establishing a conclusive positive proof of a result.

Solutions of deep and difficult problems lead to the persistent public impression that to work in mathematics, at whatever level, requires superhuman talent, and that those so endowed can understand the incomprehensible and arrive at seemingly miraculous conclusions without great effort.

As we all know, some may be more perceptive and able to see deeper results, but no one, at any stage, who does mathematics is spared much thought and greater effort. Experimentation and conjecture open the way and allow all to participate in, and benefit from, the adventure.

We need to extend public awareness to the realization that mathematics is more than arithmetic calculations, algebraic manipulations, and Euclidean geometric proofs, and that studying the calculus is not the ultimate attainment. We must counter the general perception of mathematics as a static subject, and the image of mathematicians as technicians who can solve any problem. We need to raise awareness of the vitality of mathematics as a field with a myriad of unsolved problems and numerous diverse and uncharted areas to be conquered by imaginative and creative scholars!

By adopting a problem-solving approach, we have a means of providing students with a significant mathematical experience and deeper understanding of the nature of the discipline. By introducing appropriate use of the new technology, with recognition of its limitations, we will be able to widen and enhance our students' mathematical range. By emphasizing the importance of realizing that more is required in mathematics than mere experimentation and conjecture, and by

expecting the ablest students to prove assertions and validate them, we can educate all to the full extent of their abilities. This will not only enrich their lives, but prepare them to take their place in, and contribute to, modern society, and assure the continuation of the discipline as a major force in the world.

REFERENCES

- Bell, E. T., *The Development of Mathematics*, McGraw-Hill Book Co., New York and London, 1945.
- Devlin, Keith, *Mathematics: The New Golden Age*, Penguin Books, 1988.
- Elkies, Noam, On $A^4 + B^4 + C^4 = D^4$, *Math. Comp.* 51 #184 (1988), 825–835.
- Edwards, H.M., *Riemann's Zeta Function*, Academic Press, New York and London, 1974.
- Gardner, R. J. and Wagon, Stan, At Long Last, the Circle Has Been Squared, *Notices Amer. Math. Soc.* 36 #10 (1989), 1338–1343.
- Friel, James, *An Elementary Proof for a Famous Counting Problem*, Two Year College Mathematics Readings, (1981), 218–220.
- Gelbaum, B. and Olmsted, J., *Counterexamples in Analysis*, Holden-Day, Inc., San Francisco, 1964.
- Guy, Richard, The Strong Law of Small Numbers, *Amer. Math. Monthly*, 95 (1988), 697–711.
- Horgan, John, *The Death of Proof*, Scientific American, October, 1993.
- Jaffe, A. and Quinn, F., Theoretical Mathematics: Toward a cultural synthesis of mathematics and theoretical physics, *Bulletin Amer. Math. Soc.* 29 #1 (July, 1993), 1–13.
- Laczkovich, M., Equidecomposability and discrepancy; a solution of Tarski's circle-squaring problem, *J. Reine Angew. Math.* 404 (1990), 77–117.
- Lance, T. and Thomas, E., *Arcs with Positive Measure and Space-Filling Curve*, *Amer. Math. Monthly*, 98 (1991), 124–127.
- Lander, L. J. and Parkin, T. R., Counterexamples to Euler's conjecture on sums of like powers, *Bul. Amer. Math. Soc.* 72 (1966), 1079.
- Lehman, R. Sherman, On Liouville's function, *Math. Comp.* 14 (1960), 311–320.
- Lehmer, D. H., Comments on Number Theory, Obituary: George Polya, (1985), 584–585.
- Munkres, James, *Topology, A First Course*, Prentice Hall, Inc., Englewood Cliffs, N.J., 1975.
- Nussbaum, A. E., *Equivalence by Finite Decomposition*, Columbia University Master's Thesis, 1949.
- Peterson, Ivars, *Islands of Truth—a mathematical mystery cruise*, W. H. Freeman and Co., New York, 1990.
- Peterson, Ivars, *The Mathematical Tourist*, W. H. Freeman and Co, New York, 1988.
- Rickart, C., *Structuralism in Teaching and Learning*, Mathematics Education Colloquium, Teachers College, Columbia, 1993.
- Stromberg, Karl, The Banach-Tarski Paradox, *Amer. Math. Monthly* 86 (1979) 151–161
- Tanaka, Minoru, A numerical investigation on cumulative sum of the Liouville function, *Tokyo J. Math.*, 3 (1980), 187–189.
- Wagon, Stan, *The Banach-Tarski Paradox*, Cambridge University Press. N.Y., 1985.
- Wu, Hung-Hsi, *Review of College Preparatory Mathematics 9CPB at Berkeley High School*, 1992.
- Wu., Hung-Hsi, *The Role of Open-ended Problems in Mathematics Education*, 1993.

Department of Mathematics & Computer Science
University of Missouri-St. Louis
8001 Natural Bridge Road
St. Louis, MO 63121-4499

Pebbling a Chessboard

Fan Chung, Ron Graham, John Morrison,
and Andrew Odlyzko

1. INTRODUCTION. The following puzzle has attracted some attention recently. We first learned of it through Martin Gardner [6]. A version of it appeared in *Omni* magazine in 1993 [11]. However, it was proposed over 10 years ago by Kontsevich [9], and a partial analysis of it was published shortly thereafter by Khodulev [8]. We begin with an infinite “chessboard” B covering the first quadrant. The cells of the board are labelled by integer coordinates (i, j) with $i, j \geq 0$. Initially, a single “pebble” is located in cell $(0, 0)$ (the lower left corner; see Figure 1). The first step or “move” consists of replacing this pebble by two pebbles, located at cells $(1, 0)$ and $(0, 1)$, respectively. In general, a move will consist of removing some pebble, say in cell (i, j) , and placing *two* pebbles on the board, in positions $(i + 1, j)$ and $(i, j + 1)$, *provided each of these positions is not already occupied*.

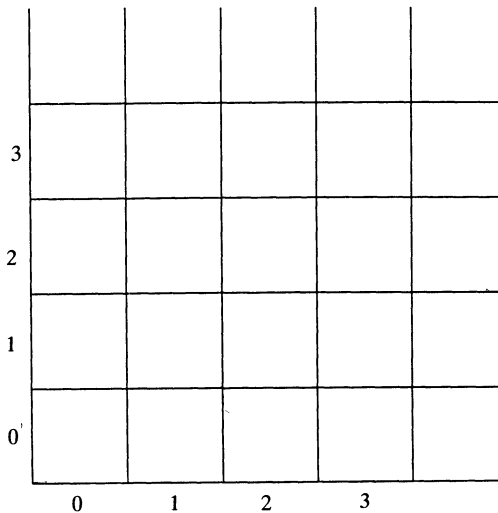


Figure 1. The starting configuration on the board B .

After k steps the board will have $k + 1$ pebbles on it. We call such configurations of pebbles *reachable configurations*. We will denote by $R(k)$ the set of reachable configurations with k pebbles, and we set $R := \bigcup_{k \geq 1} R(k)$. In Figure 2, we show the eight possible reachable configurations with at most four pebbles.

A little experimentation convinces one that in any reachable configuration, some pebble must occupy a cell having coordinates (i, j) with $i + j \leq 3$. This fact first seems to have been noted by M. Kontsevich [9]. We give the “book” proof of this in the next section. If $L(k)$ denotes the set (or “level”) $\{(i, j): i + j = k\}$ then

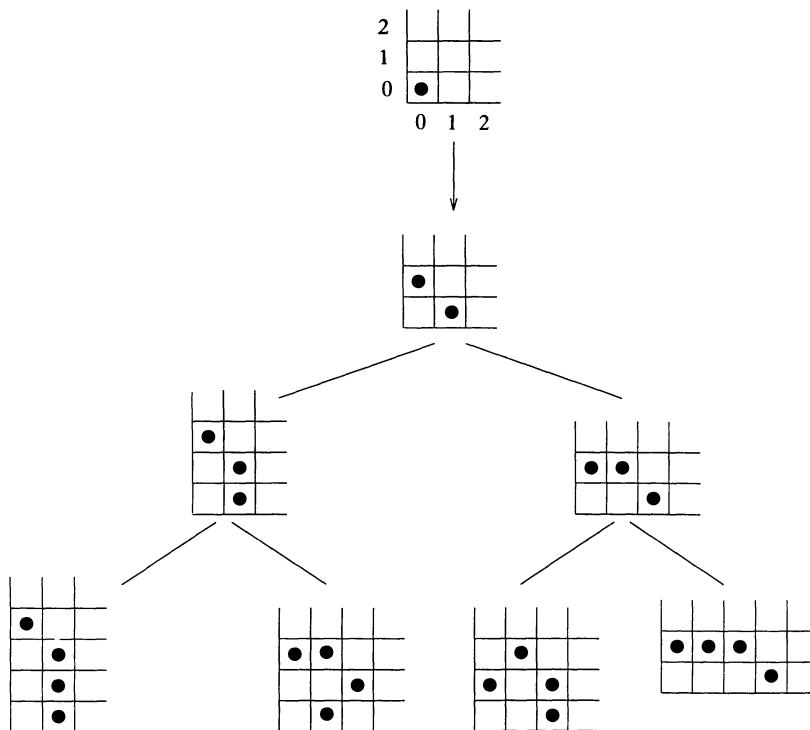


Figure 2. Reachable configurations with at most four pebbles.

we can express the above assertion by saying that $L(1) \cup L(2) \cup L(3)$ is *unavoidable*, i.e., any reachable configuration must always have some pebble in a cell in $L(1) \cup L(2) \cup L(3)$. In general, an unavoidable set is one which intersects every reachable configuration. Of course if S is unavoidable and $T \supseteq S$ then T is unavoidable. Let us call S a *minimal unavoidable* set if S is unavoidable but no proper subset of S is, and let $M(k)$ denote the family of minimal unavoidable sets with k cells.

In this note we will characterize the elements of $M(k)$ and give a polynomial time algorithm for recognizing such elements. Many of these results were first proved by Khodulev [8], and we present them here for completeness, since the paper [8] is not widely available and contains only sketches of proofs. We will also determine the asymptotic growth rates of $r(k) := |R(k)|$ and $m(k) := |M(k)|$, the sizes of $R(k)$ and $M(k)$, respectively, as $k \rightarrow \infty$. (These results are all new.) It turns out that the analysis of $r(k)$ and $m(k)$ leads to some interesting problems in asymptotic enumeration.

Further results on this problem, including generalizations to arbitrary partially ordered sets, have recently been obtained by Eriksson [4].

2. PROPERTIES OF UNAVOIDABLE SETS

Lemma 1. [9] *The set $L(1) \cup L(2) \cup L(3)$ of all (i, j) with $i + j \leq 3$ is unavoidable.*

Proof: To each cell (i, j) assign the weight $2^{-(i+j)}$. Observe that:

- (i) The total weight covered by pebbles in any reachable configuration is 1. This is so since the starting cell $(0, 0)$ has weight 1, and a move does not

change the weight of cells covered, i.e.,

$$2^{-(i+j)} = 2^{-((i+1)+j)} + 2^{-(i+(j+1))}.$$

- (ii) The total weight of *all* cells in the board is $\sum_{i,j \geq 0} 2^{-(i+j)} = 4$.
- (iii) The total weight of $L(1) \cup L(2) \cup L(3)$ is $13/4$. Thus, the weight of the *complement* of $L(3)$ is only $3/4$, and since that is less than 1, cannot contain all the pebbles of a reachable configuration. Thus, $L(1) \cup L(2) \cup L(3)$ is unavoidable. ■

However, $L(1) \cup L(2) \cup L(3)$ is not a minimal unavoidable set. The following result was proved by Khodulev [8]. It was independently conjectured by Martin Gardner [6]. The proof given here is due to Harold Reiter [14].

Lemma 2. $L(1) \cup L(2)$ is unavoidable.

Proof: As before, assign the weight $2^{-(i+j)}$ to the cell (i, j) . Observe now that any reachable configuration C has exactly one pebble on each of the boundaries $\{(i, 0): i \geq 0\}$ and $\{(0, j): j \geq 0\}$. Thus, the total weight which C can cover outside of $L(1) \cup L(2)$ is

$$2 \cdot 2^{-3} + \sum_{\substack{i,j \geq 1 \\ i+j \geq 3}} 2^{-(i+j)} = 1.$$

This implies that if C is to avoid $L(1) \cup L(2)$, it must cover *all* these cells, which is impossible since C is finite. ■

However, $L(1) \cup L(2)$ is not minimal either, as we will see later.

We should observe that for any reachable configuration C , the *set* of moves needed for reaching C is unique. Only the *order* in which these moves are executed can vary in the different ways of reaching C .

Suppose now that we relax the rules for moves by allowing the replacement of a pebble at (i, j) by pebbles at $(i + 1, j)$ and $(i, j + 1)$ even when these positions might already be occupied by pebbles. In other words, we allow the accumulation of multiple pebbles in cells during the process of reaching C . It might be helpful for this model to imagine that the pebbles first move onto the vertices of an infinite binary tree rooted at $(0, 0)$. Then the 2^k vertices in the k th level of the tree are identified in the obvious way with the $k + 1$ cells in the k th level $L(k) := \{(i, j): i + j = k\}$ of the board B .

An easy induction argument now establishes the following result.

Lemma 3. *If a configuration of pebbles (with at most one pebble per cell) can be reached by moves which **allow** accumulations of pebbles in cells, then in fact it can also be reached by the “standard” moves, i.e., those which do **not allow** accumulation.*

Given a set $X \subset B$, we define the set $M(X)$ of moves recursively as follows. Starting at level 0 and proceeding one level at a time by increasing levels, perform the moves required either to remove *all* pebbles from a cell in X , or to remove all but at most one of the pebbles from a cell not in X . Continue through the last level $L(h(X))$ containing a cell of X .

Theorem 1. $X \subset B$ is unavoidable if and only if after executing the moves in $M(X)$, some cell contains at least 3 pebbles.

Proof: Let $m(i, j)$ denote the number of pebbles in cell (i, j) after executing $M(X)$.

(i) Suppose that X is avoidable and $m(i, j) \geq 3$ for some (i, j) . Thus, either $m(i - 1, j + 1) \geq 2$ or $m(i + 1, j - 1) \geq 2$. Assume $m(i - 1, j + 1) \geq 2$ (the other case is similar). Hence, to reach *any* $C \in R$, we must move at least two pebbles off of (i, j) , and at least one off of $(i - 1, j + 1)$. But this will force $(i, j + 1)$ to have at least 3 pebbles, and will force $(i + 1, j)$ to have at least two. Thus, by induction, we can *never* reach an allowable configuration of pebbles (i.e., one in which no cell has more than one pebble), which is a contradiction.

(ii) Suppose $m(i, j) \leq 2$ for all $(i, j) \in B$. By the definition of $M(X)$,

$$m(i, j) \text{ is } \begin{cases} \leq 1 & \text{if } (i, j) \text{ has level } \leq h(X) \\ \leq 2 & \text{if } (i, j) \text{ has level } h(X) + 1 \\ = 0 & \text{if } (i, j) \text{ has level } > h(X) + 1. \end{cases}$$

A simple induction argument now shows that the excess pebbles can all be (eventually) moved to achieve a reachable configuration in R . Hence X is avoidable.

This completes the proof of Theorem 1. ■

Note that this result furnishes a polynomial-time algorithm for determining if X is a minimal unavoidable set.

3. RECURRENCES FOR MINIMAL UNAVOIDABLE SETS. Let $f(k)$ denote the number of minimal unavoidable sets consisting of k cells. For $j \geq 0$, define $B(j) = \bigcup_{i > j} L(i)$, the set of cells in levels exceeding j . Finally, for $t \geq 0$, define $f_t(k)$ to be the number of minimal unavoidable sets with k cells (i.e., of size k) in $B(t)$ where we start with the (multiple) pebble distribution of $1, \overbrace{2, 2, \dots, 2}^t, 1$ in $L(t + 1)$, and 0 in all $L(s)$, $s > t + 1$. As a convention, we take $f_t(k) = 0$ for all $k \leq 0$. Thus, $f(k) = f_0(k - 1)$ (since $(0, 0)$ must be unoccupied), and $f(k) = 0$ for $k \leq 4$. We list a set of recurrences which suffice to determine all values of $f_t(k)$:

- (i) $f_0(k) = 2f_0(k - 1) + f_1(k - 2)$;
- (ii) $f_1(k) = f_0(k) + 3f_1(k - 1) + f_2(k - 2) + 4\delta(k, 2)$ where

$$\delta(i, j) = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise;} \end{cases}$$

- (iii) For $t \geq 2$, $f_t(k) = f_{t-1}(k) + 2f_t(k - 1) + f_{t+1}(k - 2) + 2\delta(k, 1)\delta(t, 2)$.

To see why these are valid, consider (i). In Figure 3(a) we have the starting configuration for $f_0(k)$. We consider the various possibilities as to whether or not various cells in $L(1)$ are in a hypothetical minimal unavoidable set X of size k . If $(1, 0) \in X$ but $(0, 1) \notin X$ then Figure 3(b) applies and X will consist of $(1, 0)$ together with a minimal unavoidable set of size $k - 1$ arising from the two pebbles at $(2, 0)$ and $(1, 1)$. By definition, there are $f_0(k - 1)$ of these. The same argument applies if $(1, 0) \notin X$, $(0, 1) \in X$ (Figure 3(c)). On the other hand, if $(0, 1) \in X$ and $(1, 0) \in X$ then Figure 3(d) applies, and $f_1(k - 2)$ counts the number of ways of completing X . Thus, we have (i).

The other recurrences (ii) and (iii) are explained in similar ways. In Table 1, we list some of the small values of $f_t(k)$.

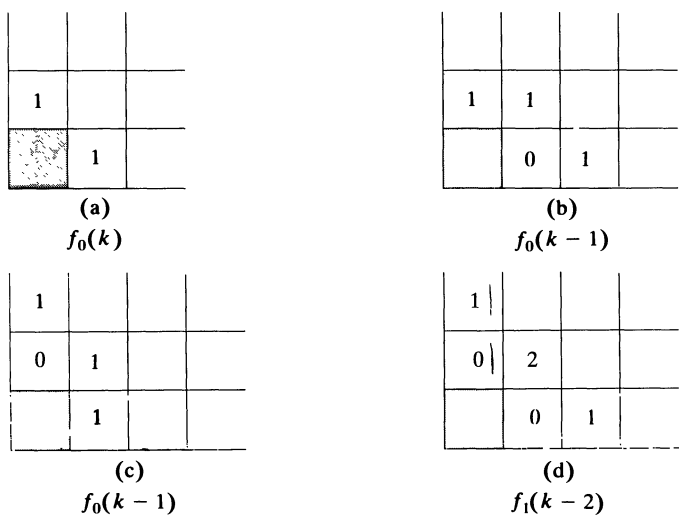


Figure 3.

TABLE 1. Values of $f_i(k)$.

10	0	2	40	464							
9	0	2	36	382							
8	0	2	32	308	2322						
7	0	2	28	242	1670	10114					
6	0	2	24	184	1154	6466					
5	0	2	20	134	758	3916					
4	0	2	16	92	466	2216	10162				
3	0	2	12	58	262	1150	4972	21296			
2	0	2	8	32	130	534	2206	9136	37872		
1	0	0	4	14	54	216	876	3574	14628	59994	
0	0	0	0	0	4	22	98	412	1700	6974	28576
	0	1	2	3	4	5	6	7	8	9	10
	k										

4. ASYMPTOTICS OF NUMBER OF MINIMAL UNAVOIDABLE SETS. The recurrences of the last section are sufficient to determine values of $f(k)$ for small k . For large values of k , we can obtain asymptotics of $f(k)$ in a simple form:

$$f(k) \sim c\gamma^{k-1} \text{ as } k \rightarrow \infty,$$

where $\gamma = 4.147899\dots$ and $c = 0.01676\dots$. More precise estimates (including definitions of γ and c) are stated at the end of this section. Since γ and c are algebraic numbers of degree 3, this estimate also shows that there is no simple expression for $f(k)$.

The derivation of the asymptotic expansion of $f(k)$ starts with the recurrences of Section 3, and proceeds through two steps. The first step is to derive an explicit expression for the generating function of $f(k)$, and the second is to obtain the asymptotics of the coefficients of that function. The second step is routine, and is sketched only briefly. The first step is the interesting one, since it involves complicated-looking functional relations that yield a surprising answer.

In order to analyze the asymptotic behavior of $f(k)$, it is convenient to introduce several auxiliary functions. The definitions are not obvious, and came from experimenting with the recurrences to find out which functions give the best

results. First, define the function $s(\cdot, \cdot)$ by

$$s(i + j, j) := f_j(i), \quad i, j \geq 0. \quad (1)$$

Next, define the generating function

$$S_i(y) := \sum_{j=0}^i s(i, j) y^j. \quad (2)$$

Thus, for example, $S_3(y) = 4y + 2y^2$. For $i \geq 3$, recurrence (iii) of Section 3 is easily seen to be equivalent to the relation

$$S_{i+1}(y) = \frac{(1+y)^2}{y} S_i(y) - \frac{1}{y} s(i, 0) + y s(i, 1). \quad (3)$$

Finally, set

$$S(x, y) := \sum_{i \geq 3} S_i(y) x^i. \quad (4)$$

Note that we are only interested in

$$\begin{aligned} \sum_{k=5}^{\infty} f(k) x^{k-1} &= \sum_{k=5}^{\infty} f_0(k-1) x^{k-1} = \sum_{i=4}^{\infty} s(i, 0) x^i \\ &= \sum_{i=4}^{\infty} S_i(0) x^i = S(x, 0). \end{aligned}$$

The additional variable y is brought in only in order to exploit the structure of recurrences for the $f_i(k)$. From (3) and (i), (ii), (iii) we obtain

$$\begin{aligned} S(x, y) &= \sum_{i \geq 3} S_i(y) x^i \\ &= x^3(4y + 2y^2) + \frac{x(1+y)^2}{y} \sum_{i \geq 3} S_i(y) x^i \\ &\quad - \frac{x}{y} S(x, 0) + xy \frac{\partial S(x, y)}{\partial y} \Big|_{y=0}. \end{aligned} \quad (5)$$

Hence

$$(y - x(1+y)^2) S(x, y) = x^3(4y^2 + 2y^3) - x S(x, 0) + xy^2 \frac{\partial S(x, y)}{\partial y} \Big|_{y=0}. \quad (6)$$

This is a complicated partial differential equation that at first sight might seem intractable. However, it can be solved explicitly. Differentiating (6) with respect to y and then setting $y = 0$, we have

$$(1 - 2x) S(x, 0) - x \frac{\partial S(x, y)}{\partial y} \Big|_{y=0} = 0. \quad (7)$$

Therefore, we can eliminate $\frac{\partial S(x, y)}{\partial y} \Big|_{y=0}$ to obtain

$$(y - x(1+y)^2) S(x, y) = (y^2(1 - 2x) - x) S(x, 0) + x^3(4y^2 + 2y^3). \quad (8)$$

On the curve

$$y = x(1+y)^2, \quad (9)$$

the coefficient of $S(x, y)$ in (8) vanishes and we have

$$S(x, 0) = x^3(4y^2 + 2y^3)/(x - y^2(1 - 2x)). \quad (10)$$

Eq. (9) implies that

$$y = (1 - 2x - (1 - 4x)^{1/2})/(2x)$$

for $|x| < 1/4$, and substituting this into (10) gives an explicit representation of $S(x, 0)$ as an algebraic function of x for $|x| < 1/4$,

$$S(x, 0) = x^2 \frac{(1 - 4x)^{1/2}(1 - 3x + x^2) - 1 + 5x - x^2 - 6x^3}{1 - 7x + 14x^2 - 9x^3}. \quad (11)$$

(Through (8) this also gives an explicit representation of $S(x, y)$ for (x, y) in a neighborhood of $(0, 0)$, but we do not need this, since $S(x, 0)$ is all that is needed to derive the asymptotics of $f(k)$.)

The final part of our analysis is now straightforward. The explicit form of $S(x, 0)$ shows that $S(x, 0)$ is analytic in $|x| < 1/4$ except at zeros of the denominator, i.e. at $x = 1/\gamma$, where $\gamma = 4.14789903 \dots$ satisfies

$$\gamma^3 - 7\gamma^2 + 14\gamma - 9 = 0. \quad (12)$$

Direct substitution into the formula for $S(x, 0)$ then shows that $S(x, 0)$ actually does have a simple pole at $x = 1/\gamma$, but (in view of the preceding discussion) no other singularities in $|x| < 1/4$. By the standard methods [2, 3, 7, 12], we can therefore write

$$f(k) = f_0(k - 1) = s(k - 1, 0) = [x^{k-1}]S(x, 0) = c\gamma^{k-1} + O(4.01^k),$$

where

$$c = \lim_{x \rightarrow 1/\gamma} S(x, 0)(1 - \gamma x) = 0.016762198 \dots \quad (13)$$

and satisfies (after some messy but routine computation best done with a symbolic algebra system)

$$7533c^3 + 10726c^2 + 5068c - 88 = 0. \quad (14)$$

5. THE NUMBER OF PEBBLE CONFIGURATIONS. In this section we will treat the problem of enumerating the number of distinct reachable configurations with k pebbles. We denote this number by $g(k)$. As was true for the asymptotics of $f(k)$, it is the derivation of an explicit generating function for the $g(k)$ that presents the main challenge here.

As before, let us define $g_t(k)$ to be the number of k -pebble reachable configurations where we start with the initial pebble distribution of $1, \overbrace{2, 2, \dots, 2}^t, 1$ in L_{t+1} , and 0 in all L_s , $s > t + 1$ (and we restrict ourselves to cells just in $B(t) = \bigcup_{s \geq t+1} L_s$). Thus, $g(k) = g_0(k)$ for $k \geq 2$. Arguing along the same lines as before, it is not hard to derive the following recurrences for the $g_t(k)$:

- (i') $g_0(k) = 2g_0(k - 1) + g_1(k) + \delta(k, 2)$;
- (ii') $g_1(k) = g_0(k - 3) + 2g_1(k - 2) + g_2(k - 1) + g_1(k - 4)$;
- (iii') For $t \geq 2$,

$$g_t(k) = g_{t-1}(k - t - 2) + 2g_t(k - t - 1) + g_{t+1}(k - t).$$

Now set

$$\begin{aligned} h_i(k) &:= g_i(k + i), \\ H_i(x) &:= \sum_{k=0}^{\infty} h_i(k) x^k, \\ H(x, y) &:= \sum_{i=0}^{\infty} H_i(x) y^i. \end{aligned} \quad (15)$$

TABLE 2. Values of $h_t(k)$.

t	2	0	0	0	0	0	0	0	1	2
	1	0	0	0	0	1	2	6	13	33
	0	0	0	1	2	4	9	20	46	105
		0	1	2	3	4	5	6	7	8
		k								

Some values of $h_t(k)$ are shown in Table 2. Straightforward computation using (15) and (i'), (ii'), (iii') shows

$$H(x, y) = x^2 + \left(\frac{1}{y} + 2x + x^2y \right) H(x, xy) - \frac{1}{y} H(x, 0) + x^4 y H_1(x). \quad (16)$$

Since $H_1(x) = \left. \frac{\partial H(x, y)}{\partial y} \right|_{y=0}$, we have

$$yH(x, y) = x^2 y + (1 + xy)^2 H(x, xy) - H(x, 0) + x^4 y^2 \left. \frac{\partial H(x, y)}{\partial y} \right|_{y=0}. \quad (17)$$

Differentiating (17) with respect to y , and setting $y = 0$ implies

$$H(x, 0) = x^2 + x \left. \frac{\partial H(x, y)}{\partial y} \right|_{y=0} + 2xH(x, 0). \quad (18)$$

Substituting

$$x \left. \frac{\partial H(x, y)}{\partial y} \right|_{y=0} = (1 - 2x)H(x, 0) - x^2$$

into (17) gives

$$yH(x, y) = (1 + xy)^2 H(x, xy) + (x^3 y^2 - 2x^4 y^2 - 1)H(x, 0) + x^2 y - x^5 y^2 \quad (19)$$

which is the basic relation for $H(x, y)$ we will use. This is more difficult to analyze than the corresponding functional equation (8) for $S(x, y)$ but we still can obtain significant information about its asymptotic behavior.

To begin, from (15) and (19) we have

$$\begin{aligned} (1 - 2x)H_0(x) &= xH_1(x) + x^2 \\ H_1(x) &= x^2(H_2(x) + 2H_1(x) + H_0(x)) + x^4 H_1(x) \end{aligned} \quad (20)$$

and for $n \geq 2$,

$$H_n(x) = x^{n+1}(H_{n+1}(x) + 2H_n(x) + H_{n-1}(x)).$$

Therefore,

$$\begin{aligned} H_1(x) &= \left(\frac{1 - 2x}{x} \right) H_0(x) - x, \\ H_2(x) &= \left(\frac{1 - x^4}{x^2} \right) H_1(x) - 2H_1(x) - H_0(x) \\ &= \frac{1}{x^3} (((1 - 2x^2 - x^4)(1 - 2x) - x^3)H_0(x) - x^2(1 - 2x^2 - x^4)), \end{aligned}$$

and for $n \geq 3$,

$$H_n(x) = \frac{1}{x^n}((1 - 2x^n)H_{n-1}(x) - x^n H_{n-2}(x)).$$

It then follows by induction that

$$H_n(x) = x^{-\binom{n+1}{2}}(q_n(x)H_0(x) - x^2 p_n(x)) \quad (21)$$

where

$$\begin{aligned} q_1(x) &= 1 - 2x, & p_1(x) &= 1, \\ q_2(x) &= (1 - 2x^2 - x^4)(1 - 2x) - x^3, & p_2(x) &= 1 - 2x^2 - x^4 \end{aligned}$$

and for $n \geq 3$,

$$\begin{aligned} q_n(x) &= (1 - 2x^n)q_{n-1}(x) - x^{2n-1}q_{n-2}(x), \\ p_n(x) &= (1 - 2x^n)p_{n-1}(x) - x^{2n-1}p_{n-2}(x) \end{aligned} \quad (22)$$

(where we can consider (21) as a formal power series identity). From (22) we see that

$$\begin{aligned} q(x) &= \lim_{n \rightarrow \infty} q_n(x), \\ p(x) &= \lim_{n \rightarrow \infty} p_n(x) \end{aligned}$$

exist as formal power series and that

$$[x^k]q(x) = [x^k]q_n(x), \quad [x^k]p(x) = [x^k]p_n(x)$$

for $n \geq k + 1$. Note that by (21), increasing powers of x divide $q_n(x)H_0(x) - x^2 p_n(x)$ as $n \rightarrow \infty$. Thus, we have

$$H_0(x) = \frac{x^2 p(x)}{q(x)} \quad (23)$$

as a formal power series.

From (22) and (23) it now follows (cf. [5]) that $H_0(x)$ can be written as the continued fraction

$$H_0(x) = \cfrac{x^2}{1 - 2x - \cfrac{x^3}{1 - 2x - x^4 - \cfrac{x^5}{1 - 2x^3 - \cfrac{x^7}{1 - 2x^4 - \cfrac{x^9}{1 - 2x^5 - \cfrac{x^{11}}{1 - 2x^6 - \dots}}}}}} \quad (24)$$

Although this continued fraction is similar to some studied by Ramanujan (see [1], [15]), it does not seem to have appeared in the literature before.

The recurrences (22) imply that $p(x)$ and $q(x)$ are analytic in the disc $\{x: |x| < 1\}$, and so $H_0(x)$ is meromorphic for $|x| < 1$. To determine the asymptotic behavior of $H_0(x)$, we need to look at the zeros of $q(x)$. It turns out that in the disc $\{x: |x| \leq 1/2\}$, $q(x)$ has only a simple zero at $\beta_1 = 1/\alpha$ where $\alpha = 2.321642199494 \dots$. This implies

$$h_0(k) = [x^k]H_0(x) = c_1 \alpha^k + O(2^k) \quad (25)$$

where

$$c_1 = \frac{-\beta_1 p(\beta_1)}{q'(\beta_1)}.$$

In fact, $q(x)$ has zeros of multiplicity one at

$$\beta_1 = 0.430729593 \dots$$

$$\beta_2 = 0.685754744 \dots$$

$$\beta_3 = -0.704352541 \dots$$

$$\beta_4 = 0.782572917 \dots$$

and no other zeros in $\{x: |x| \leq 0.8\}$. A more careful analysis shows that (25) can be improved to

$$h_0(k) = \sum_{j=1}^4 \frac{-p(\beta_j)}{q'(\beta_j)} \beta_j^{-k+1} + O((5/4)^k), \quad (26)$$

and even better approximations can be obtained with more effort.

The basic technique for proving (25) is given, for example, in [13]. We give a brief sketch here. To begin, computation shows that $q(x)$ starts as follows:

$$\begin{aligned} q(x) = & 1 - 2x - 2x^2 + x^3 + x^4 + 7x^5 \\ & + 2x^6 + 5x^7 - 4x^8 - 7x^9 - 9x^{10} - 14x^{11} - \dots \end{aligned}$$

Let

$$Q(x) = 1 - 2x - 2x^2 + x^3 + x^4 + 7x^5 + \dots + 46x^{19} \quad (27)$$

consist of the first 20 terms of $q(x)$. It is not hard to verify that $|Q(x)| \geq 1/20$ for $|x| = 1/2$. We want to show that $Q_1(x) = q(x) - Q(x)$ is small on $|x| > 1/2$. For $|x| = 3/4$, computation shows that $|q_5(x)| \leq 10$, $|q_6(x)| \leq 10$. By (22),

$$|q_n(x)| \leq \left(1 + 2\left(\frac{3}{4}\right)^n\right) |q_{n-1}(x)| + \left(\frac{3}{4}\right)^{2n-1} |q_{n-2}(x)|.$$

Therefore,

$$|q(x)| \leq 30 \quad \text{for } |x| = 3/4, \quad (28)$$

which implies for $|y| = 1/2$,

$$\begin{aligned} |Q_1(y)| &\leq \sum_{m=20}^{\infty} |[x^m]q(x)| |y|^m \\ &\leq 30 \sum_{m=20}^{\infty} (2/3)^m \leq 90 \left(\frac{2}{3}\right)^{20}. \end{aligned}$$

Thus, $|Q_1(y)| < |Q(y)|$ for $|y| = 1/2$, so by Rouché's theorem, $Q(y)$ and $q(y)$ have the same number of zeros in $\{y: |y| \leq 1/2\}$. However, direct computation shows that $Q(y)$ has exactly one zero in this region, and therefore, so does $q(y)$. Consequently, $\beta_1 = 0.430729593 \dots$ is the only zero of $q(y)$ in $\{y: |y| \leq 1/2\}$.

The recurrence (22) also gives an effective method for computing the other zeros β_j , as well as the values of $p(\beta_j)$, $q'(\beta_j)$ and $c_1 = 0.12268707 \dots$.

It seems unlikely that there is as simple an expression for $g(k)$ as the one we have for $f(k)$. Poles of continued fractions such as that of (24) can seldom be

expressed in closed form, and are expected to be usually transcendental. There are few rigorous results or methods. On the other hand, accurate numerical approximations are almost always easy to obtain.

6. SOME HISTORY AND ACKNOWLEDGMENTS. It seems [10] that the original problem of showing that $\bigcup_{i=0}^4 L(i)$ is unavoidable appeared as Question 5 in the Spring 1981 Senior Paper of the Tournament of the Towns (in the former Soviet Union) where it is attributed to M. Kontsevich. The solution was presented at the first World Federation of National Mathematics Competitions Conference held at the Univ. of Waterloo in 1990. (We are indebted to Andy Liu for this bit of scholarship.)

We would particularly like to thank Martin Gardner for once again bringing to our attention a beautiful problem which looks deceptively simple and yet offers interesting challenges. We thank M. Kontsevich for informing us of the references [9, 8]. We are also grateful to H. Eriksson and D. E. Knuth for their comments and corrections to an earlier draft.

REFERENCES

1. G. Andrews, An introduction to Ramanujan's "lost notebook", *Amer. Math. Monthly* 86 (1979), 89–108.
2. E. A. Bender, Asymptotic methods in enumeration, *SIAM Review* 16 (1974), 485–515.
3. E. A. Bender and S. G. Williamson, *Foundations of Applied Combinatorics*, Addison-Wesley, 1991.
4. H. Eriksson, Pebblings, to be published.
5. P. Flajolet, Combinatorial aspects of continued fractions, *Discrete Math.* 32 (1980), 125–161.
6. Martin Gardner (personal communication).
7. R. L. Graham, D. E. Knuth, and O. Patashnik, *Concrete Mathematics*, Addison-Wesley, 1989.
8. A. Khodulev, Pebble spreading, *Kvant.* July 1982, pp. 28–31 and 55. (In Russian.)
9. M. Kontsevich, Problem M715, *Kvant.* Nov. 1981, p. 21. (In Russian.)
10. Andy Liu (personal communication).
11. S. Morris, Counter-intuitive puzzle, *Omni*, April 1993. (Solution by N. Konstantinov in the August 1993 issue.)
12. A. M. Odlyzko, Asymptotic enumeration methods, in *Handbook of Combinatorics*, R. L. Graham, M. Grötschel, and L. Lovász, eds., North-Holland, 1995, to appear.
13. A. M. Odlyzko and A. S. Wilf, The editor's corner: n coins in a fountain, *Amer. Math. Monthly* 95 (1988), 840–843.
14. Harold Reiter (personal communication).
15. G. Szekeres, A combinatorial interpretation of Ramanujan's continued fractions, *Canad. Math. Bull.* 11 (1968), 405–408.

Chung:
Bell Communications Research
Morristown, NJ 07960
Chung: frkc@bellcore.com

Graham, Morrison, and Odlyzko:
AT & T Bell Laboratories
Murray Hill, NJ 07974
Graham: rlg@research.att.com
Morrison: jam@research.att.com
Odlyzko: amo@research.att.com

Drums That Sound the Same

S. J. Chapman

1. INTRODUCTION. In 1966 Kac [7] asked the question “Can you hear the shape of a drum?”, that is, if you know the frequencies at which a drum vibrates, can you determine its shape? Mathematically this corresponds to the following problem. If u is the displacement of a membrane D from its mean position, then u satisfies

$$\nabla^2 u = \frac{\partial^2 u}{\partial t^2}, \text{ in } D, \quad (1)$$

$$u = 0, \text{ on } \partial D. \quad (2)$$

Seeking a solution by separation of variables $u(x, y, t) = \psi(t)\phi(x, y)$ yields

$$\frac{\phi_{xx} + \phi_{yy}}{\phi} = \frac{\psi_{tt}}{\psi} = \text{constant} = \lambda, \text{ say.}$$

Hence

$$u = \sin(\sqrt{\lambda} t)\phi(x, y), \quad (3)$$

where

$$\nabla^2 \phi + \lambda \phi = 0, \text{ in } D, \quad (4)$$

$$\phi = 0, \text{ on } \partial D. \quad (5)$$

This is an eigenvalue problem: there exists a nonzero solution ϕ only for certain values of λ known as eigenvalues. The set of eigenvalues is known as the

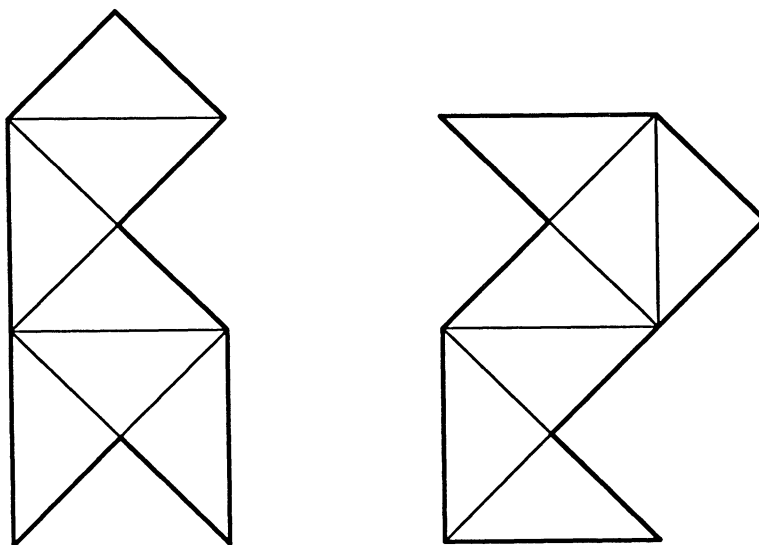


Figure 1

eigenvalue spectrum, and is discrete in this case. We see by equation (3) that the eigenvalues λ are the squares of the frequencies of vibration, and that each eigensolution can be viewed as a standing wave on the domain D . The general solution of (1), (2) is a superposition of these special solutions.

Kac's question is now the following: are two domains with the same eigenvalue spectrum (where the eigenvalues are counted with multiplicities) necessarily congruent?

It has been shown that the eigenvalues do determine certain properties of D , for example the area, the circumference, and the number of connected components [7]. However, the answer to the question is in fact no. Figure 1 shows an example of two domains with exactly the same eigenvalue spectrum which was given by Gordon et al. [5] (see also [6]).

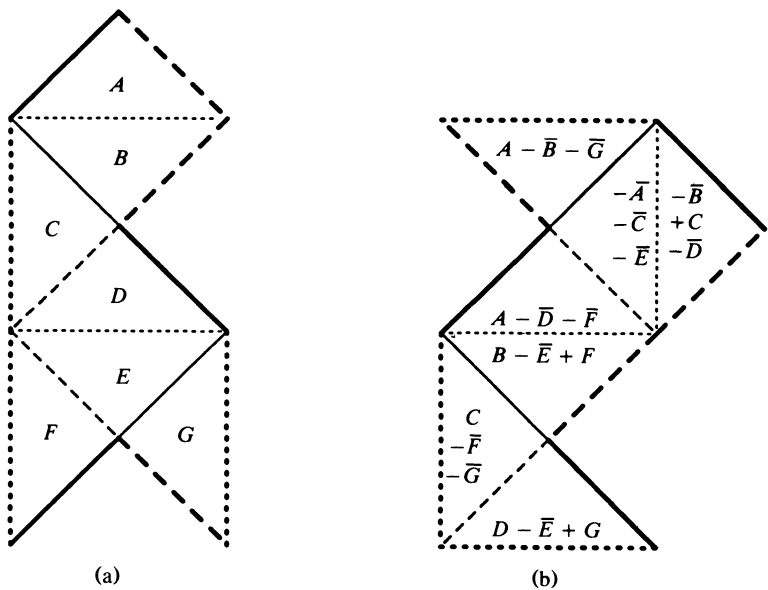


Figure 2

A simple proof that the eigenvalues are identical has been given by Berard [3] (see also [1, 2]), who constructs the map shown in Figure 2, which takes an eigenfunction for the first domain and maps it onto an eigenfunction for the second domain, with the same eigenvalue λ . Here $A + B$ means that to obtain the value of the function in that triangle we add the values of the function at the corresponding points in triangles A and B . We have used different types of lines for the edges of the triangles to help make it clear how each should be orientated when making this identification. In some cases it is necessary to reflect the triangle about its line of symmetry, and this we have indicated by \bar{A} . Only the zero function maps to the zero function, which implies that for any eigenfunction of the first eigenvalue problem there is a corresponding eigenfunction of the second eigenvalue problem, with the same eigenvalue λ . Thus any eigenvalue of the first problem is also an eigenvalue of the second problem (including multiplicities). A similar map of an eigenfunction of the second problem to one of the first shows that any eigenvalue of the second problem is also an eigenvalue of the first, and therefore the two domains are isospectral.

Here we give an interpretation of the transposition of a solution of the first problem to one of the second problem in terms of paper folding. This will allow us to generate many new isospectral domains, including a simple example in which the eigenvalues can be calculated explicitly. We note that the method of transposition has also recently been used by Buser et al. [4] to generate new examples of isospectral plane domains.

2. PAPER FOLDING. Consider a paper cutout of a domain, and a function which is zero on the boundary of the domain. We now fold the paper to create a new domain. We define a function, the transposition, on the new domain, by adding the values of the original function at points that lie on top of one another, with the convention that if the paper is reversed then the function is subtracted rather than added. This function will automatically be zero on the boundary of the new domain, since it will be zero along any fold of the paper, as well as on any edge. The reader may find it helpful in what follows to actually construct the shapes by folding paper.

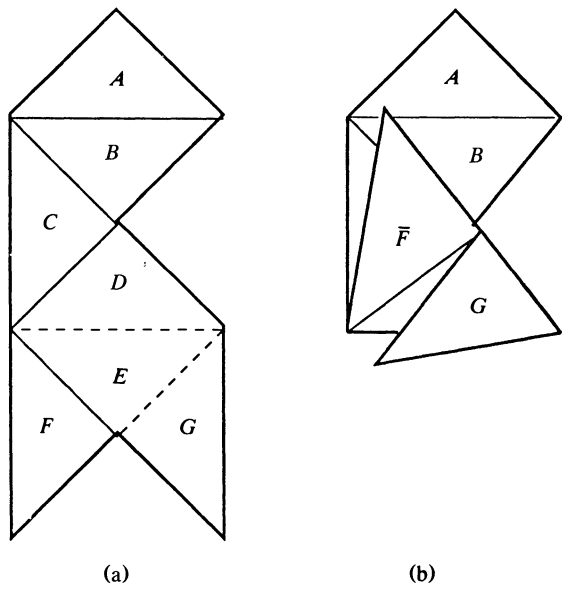


Figure 3

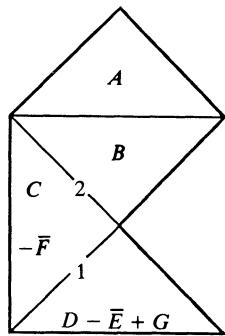


Figure 4

For example consider the domain shown in Figure 2a, and a function on this domain which is zero on the boundary. We label the triangles A to G on the front and \bar{A} to \bar{G} on the back and fold the domain as shown in Figure 3 (where a dotted line indicates a fold of the paper). We then obtain a function on the domain shown in Figure 4 which is zero on the boundary.

We now take several copies of the original domain D and fold them to create domains D_1, D_2 , etc. We glue these together to create a new domain D^* , and define the transposition on this domain to be the sum of the transpositions on D_1, D_2 , etc. Now, if we can glue the domains D_1, D_2 , etc. together in such a way that the first derivative of the transposition is continuous, then we will have actually created an eigenfunction on the new domain D^* .¹

In order for the first derivative of the transposition to be continuous it is sufficient that

- (1) Every fold lies along an outside edge of the new shape.
- (2) Each edge of each copy of the original shape that lies in the interior of the final shape must be adjacent to its reflection on an associated copy of the original shape.

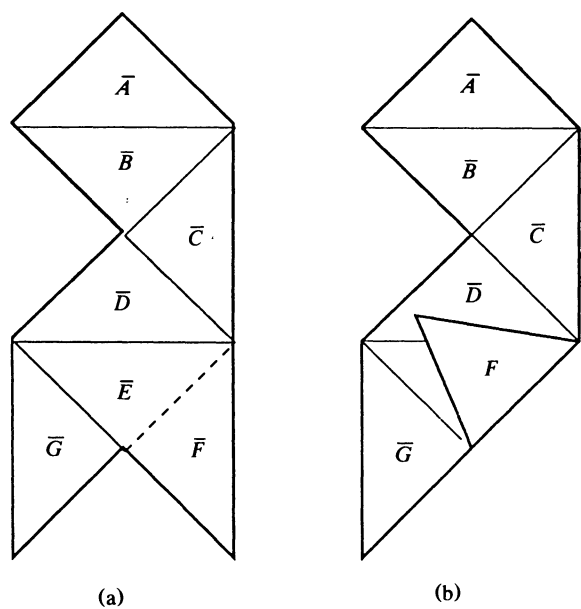


Figure 5

For example, the first derivative of the transposition is discontinuous across the lines (1) and (2) in Figure 4. However, if we add the same initial drum shape folded as shown in Figure 5 then we ensure continuity of the first derivative across the lines (1) and (2) (Figure 6), though the first derivative is now discontinuous

¹The resulting function is once differentiable and satisfies the eigenequation except possibly on the seams. Such a function is a weak solution of the equation and therefore by elliptic regularity a strong solution; see G. Folland, Introduction to PDE, PUP, 1976: specifically apply Corollary 6.28 repeatedly and then Corollary 6.10 to find that the transposed function is indeed an eigenfunction.

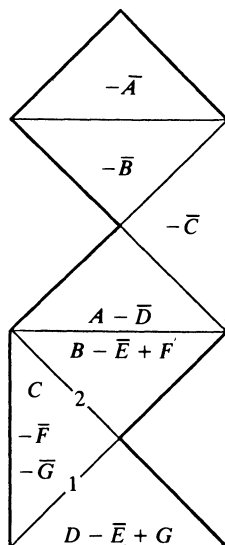


Figure 6

across other lines. Note that the continuity of the transposed function itself is automatic since the original function is zero on the boundary of each component.

Figure 7 shows how the pieces fit together for the example given in the introduction. Three copies of the original shape (Fig. 2a) are folded along the dotted lines shown in Fig. 7a, to give the shapes shown in Fig. 7c (Figure 7b shows a three dimensional view of how each piece will look before it is squashed flat). These shapes are then superimposed to create the shape shown in Fig. 2b (the dotted lines in Fig. 7c indicate the position of each component in the new shape).

Another example of isospectral domains, and the transposition of a solution on one domain onto a solution on the other domain, is given in Figures 8 and 9. The cuts in these figures are to be interpreted as having zero width, and are shown for clarity.

Since the method of construction of the transposition depends only on folding along the edges of the triangles, there is no need for the triangles to be right-angled. All that is important is that the two triangles adjacent to a fold lie on top of one another when the paper is folded. If we think of the shapes in Figure 2 as being constructed from a single triangle A by a series of reflections about its edges, then it is not the shape of A , but the series of reflections which is important. Choose any other triangle in place of A in Figure 10a, and perform the same series of reflections to obtain a new shape. Now place the same triangle in position d of Figure 10b, with the same orientation, and perform the series of reflections that created 10b from the basic right-angled triangle d . The two shapes obtained will then be isospectral, since the same map transposing eigenfunctions of one domain onto the other domain will work as before. For example, if we use the triangle shown in Figure 11a as our basic building block, we find that the domains shown in Figure 11b are isospectral.

It is not necessary for the triangle to be isosceles. If we take the triangle shown in Figure 12a as our basic building block, we find that the domains shown in Figure 12b are isospectral.

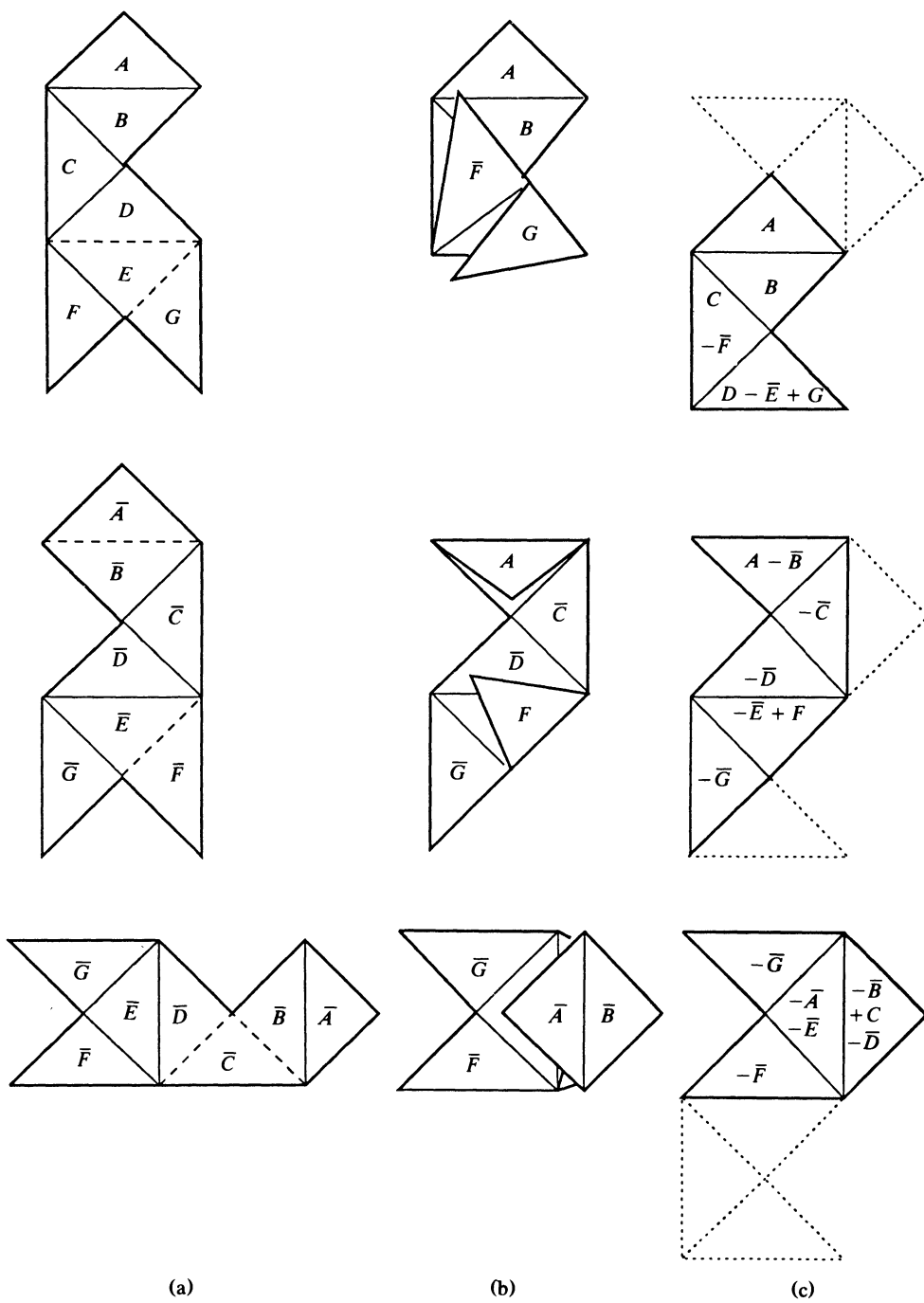


Figure 7

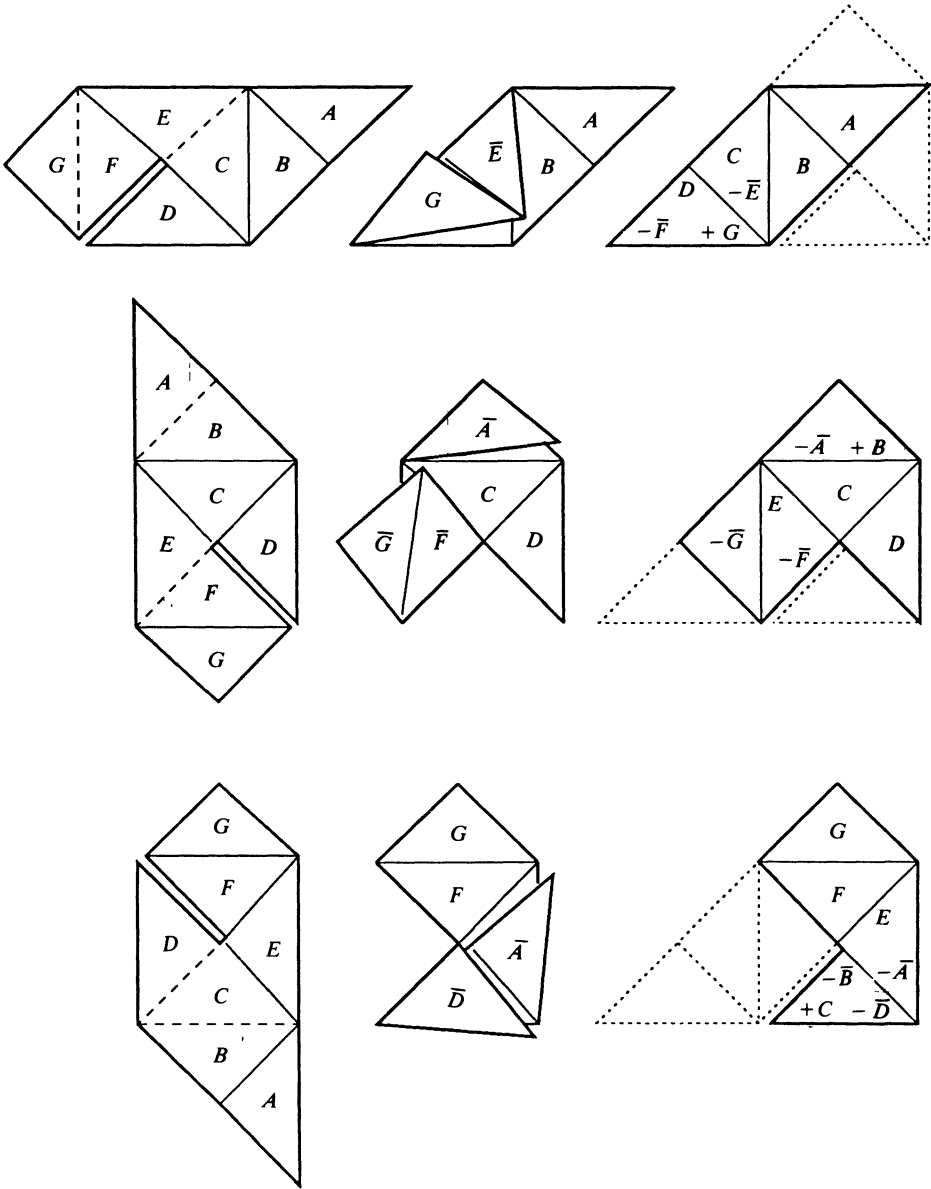


Figure 8

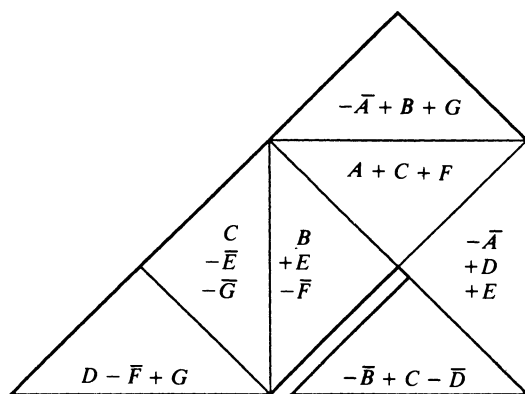


Figure 9

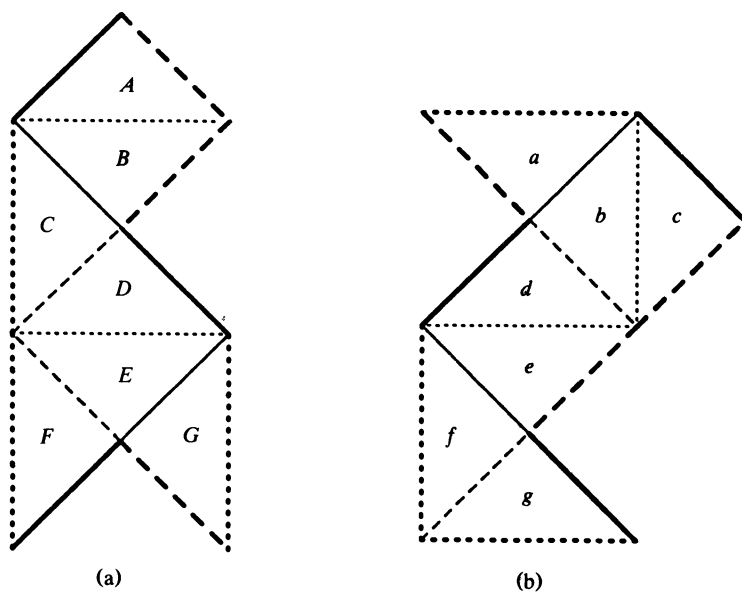


Figure 10

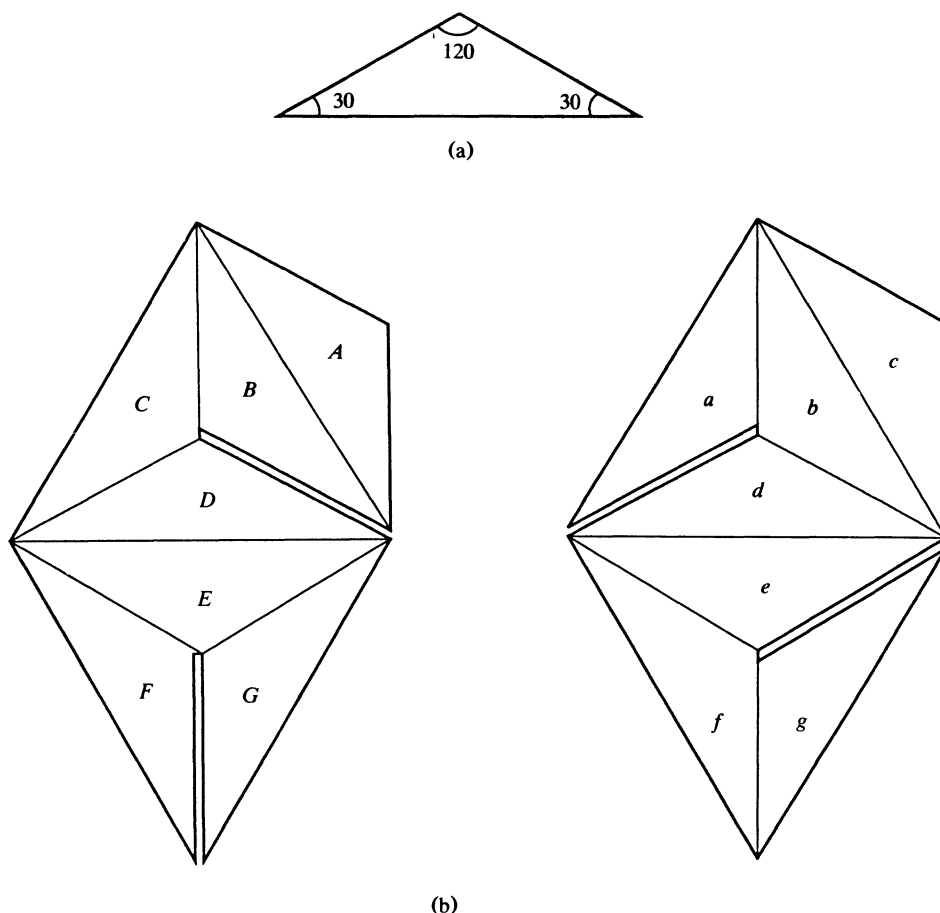


Figure 11

By the same considerations, it is not even necessary for the basic shape to be a triangle. Any shape with at least three edges will do. We simply choose three edges to represent the three sides of the triangle, about which we will reflect the shape. If we then follow the same pattern of reflection that created the original shapes of Fig. 2 from the basic right-angled triangle, then we have again isospectral drums. The example shown in Figure 13 uses squares.

In fact, the basic starting shape can be as complicated as you like. To construct different shapes, take any of the previous shapes constructed of triangles, squares etc. and fold both shapes until a single triangle (for example) remains. Place the two resulting triangles on top of one another (with the correct orientation, i.e. so that the solid, dotted and dashed lines match up). Now cut out shapes as when making paper dolls. The shapes obtained when the paper is unfolded will again have exactly the same eigenvalues as each other, since the same one-to-one correspondence between solutions will hold as before. We note that the cutout will be in one piece if and only if there is a segment left uncut on each edge.

Figure 14 shows a simple example. These isospectral domains were also discovered by Gordon et al. More exotic shapes can also be made, as shown in Figure 15.

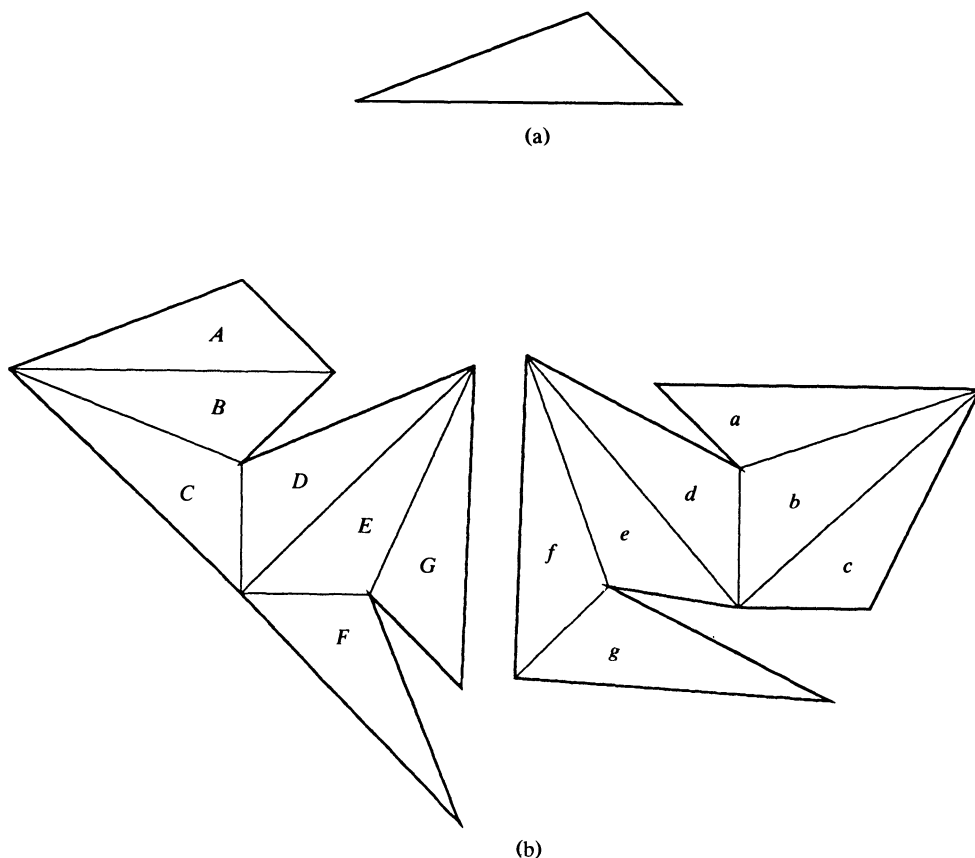


Figure 12

In this way a simpler example of drums with the same eigenvalues can be constructed—one in which the eigenvalues can actually be calculated explicitly. Consider the original example, folded and cut along one edge as shown in Figure 16a. Then the drums obtained are as shown in Figure 16b. Discarding the single small triangle, which appears once in each drum, we have the domains shown in Figure 17. The spectrum of each of the disconnected domains in Figure 17 is equal to the union of the spectra of each of the components (with the multiplicity of an eigenvalue being equal to the sum of its multiplicities in the components), since each of the components vibrates independently. The eigenfunctions for a rectangle of length a and width b are

$$\sin \frac{n\pi x}{a} \sin \frac{m\pi y}{b}, \quad n, m \text{ integers},$$

with corresponding eigenvalues $\lambda = \pi^2((n/a)^2 + (m/b)^2)$. For a right-angled isosceles triangle with short sides of length c the eigenfunctions are

$$\sin \frac{i\pi x}{c} \sin \frac{j\pi y}{c} - \sin \frac{j\pi x}{c} \sin \frac{i\pi y}{c}, \quad i, j \text{ integers}, \quad i > j,$$

with corresponding eigenvalues $\lambda = \pi^2((i/c)^2 + (j/c)^2)$. Thus we find that the eigenvalues for each domain are as shown in Fig. 17. We shall now show that the

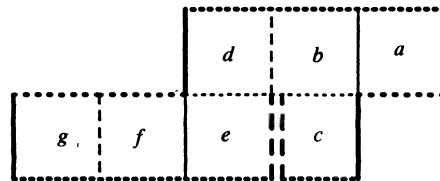
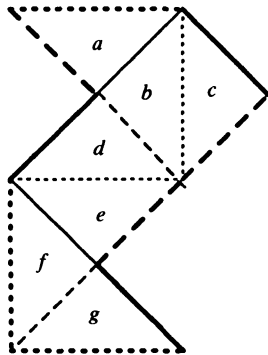
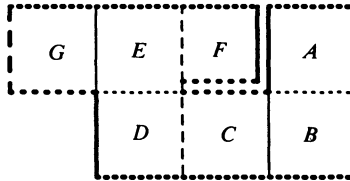
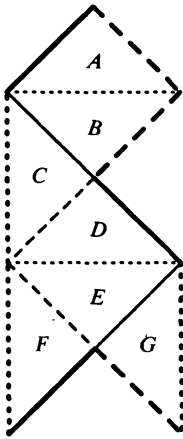


Figure 13

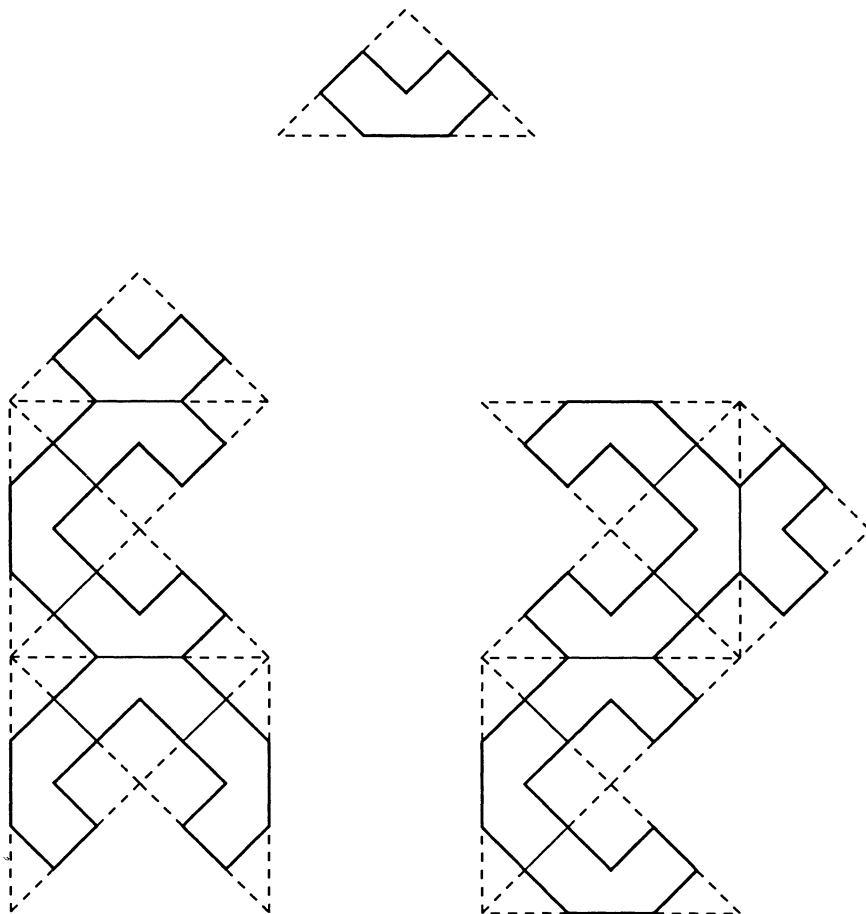


Figure 14

combined eigenvalues of the two domains of Figure 17a are identical to the combined eigenvalues of the two domains of Figure 17b.

With N even we set $N = 2n$ and $M = m$. Then we have $(N/2)^2 + M^2 = n^2 + m^2$. When N is odd we set $i = \max(N, 2M)$, $j = \min(N, 2M)$. Then $(N/2)^2 + M^2 = (i/2)^2 + (j/2)^2$, and $i > j$. This takes care of the eigenvalues in which one of i, j is even and the other is odd. If we set $i = I + J$ and $j = I - J$ then $(I^2 + J^2)/2 = (i/2)^2 + (j/2)^2$, and $i > j$. This takes care of the eigenvalues in which either i and j are both even or both odd.

Finally, we note that the same procedure works if the boundary condition

$$u = 0 \quad \text{on } \partial D,$$

is modified to

$$\frac{\partial u}{\partial n} = 0 \quad \text{on } \partial D,$$

if we modify our convention and add reflected triangles instead of subtracting them. Thus all the isospectral domains found previously are also isospectral with this new boundary condition.

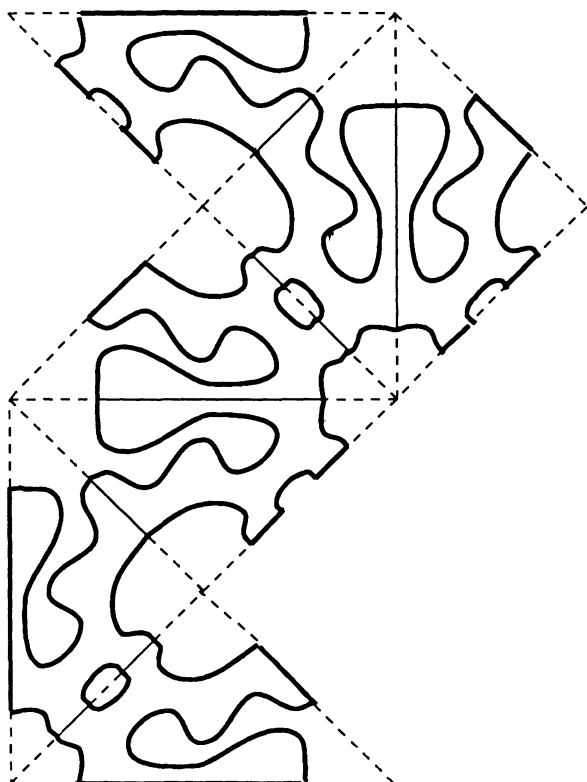
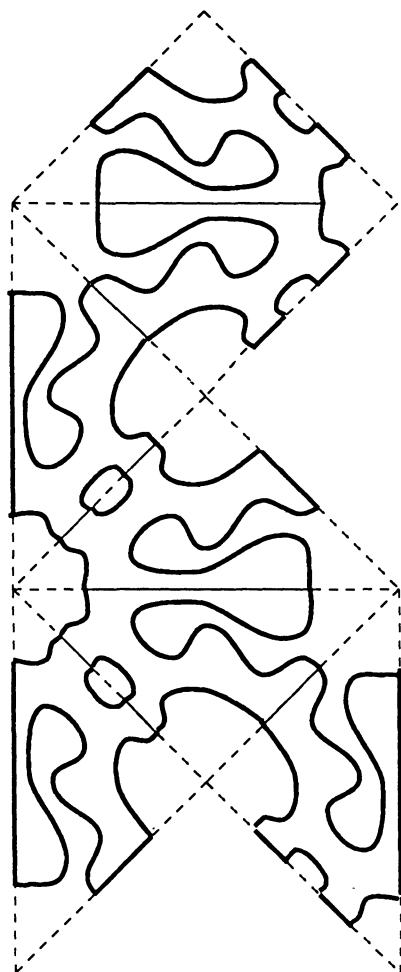


Figure 15

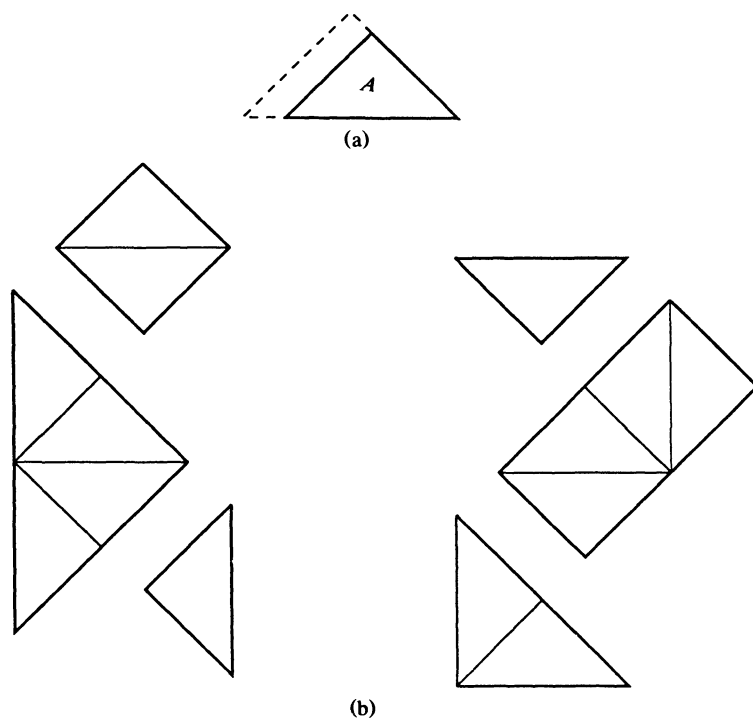


Figure 16

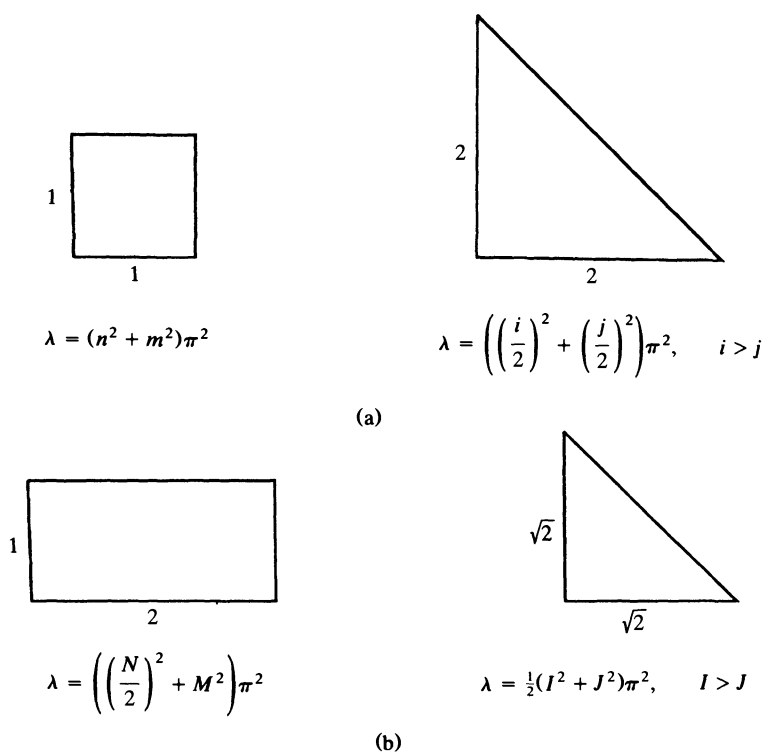


Figure 17

ACKNOWLEDGMENT. I would like to thank Prof. J. B. Keller for bringing this problem to my attention, and for useful discussions.

REFERENCES

1. Berard, P., Transplantation et Isospectralite I, *Math. Ann.*, 292, 547–559 (1992).
2. Berard, P., Transplantation et Isospectralite II, *Math. Ann.*, to appear.
3. Berard, P., Domaines plans isospectraux a la Gordon-Webb-Wolpert: une preure terra a terra, Preprint.
4. Buser, P., Conway, J., Doyle, P. and Semmler, K-D., Some planar isospectral domains, Preprint.
5. Gordon, C., Webb, D. and Wolpert, S., Isospectral plane domains and surfaces via Riemannian orbifolds, *Invent. Math.*, 110, 1–22 (1992).
6. Gordon, C., Webb, D. and Wolpert, S., You cannot hear the shape of a drum, *B. Am. Math. S.*, 27(1), 134–138 (1992).
7. Kac, M., Can one hear the shape of a drum?, *Am. Math. Monthly*, 73, 4, 1–23 (1966).

Mathematical Institute
24-29 St. Giles'
Oxford, OX1 3LB
England
jchapman@vax.ox.ac.uk

PICTURE PUZZLE (from the collection of Paul Halmos)



Half of a man and wife team
(see page 154.)

Down With Determinants!

Sheldon Axler

1. INTRODUCTION. Ask anyone why a square matrix of complex numbers has an eigenvalue, and you'll probably get the wrong answer, which goes something like this: The characteristic polynomial of the matrix—which is defined via determinants—has a root (by the fundamental theorem of algebra); this root is an eigenvalue of the matrix.

What's wrong with that answer? It depends upon determinants, that's what. Determinants are difficult, non-intuitive, and often defined without motivation. As we'll see, there is a better proof—one that is simpler, clearer, provides more insight, and avoids determinants.

This paper will show how linear algebra can be done better without determinants. Without using determinants, we will define the multiplicity of an eigenvalue and prove that the number of eigenvalues, counting multiplicities, equals the dimension of the underlying space. Without determinants, we'll define the characteristic and minimal polynomials and then prove that they behave as expected. Next, we will easily prove that every matrix is similar to a nice upper-triangular one. Turning to inner product spaces, and still without mentioning determinants, we'll have a simple proof of the finite-dimensional Spectral Theorem.

Determinants are needed in one place in the undergraduate mathematics curriculum: the change of variables formula for multi-variable integrals. Thus at the end of this paper we'll revive determinants, but not with any of the usual abstruse definitions. We'll define the determinant of a matrix to be the product of its eigenvalues (counting multiplicities). This easy-to-remember definition leads to the usual formulas for computing determinants. We'll derive the change of variables formula for multi-variable integrals in a fashion that makes the appearance of the determinant there seem natural.

A few friends who use determinants in their research have expressed unease at the title of this paper. I know that determinants play an honorable role in some areas of research, and I do not mean to belittle their importance when they are indispensable. But most mathematicians and most students of mathematics will have a clearer understanding of linear algebra if they use the determinant-free approach to the basic structure theorems.

The theorems in this paper are not new; they will already be familiar to most readers. Some of the proofs and definitions are new, although many parts of this approach have been around in bits and pieces, but without the attention they deserved. For example, at a recent annual meeting of the AMS and MAA, I looked through every linear algebra text on display. Out of over fifty linear algebra

Many people made comments that helped improve this paper. I especially thank Marilyn Brouwer, William Brown, Jonathan Hall, Paul Halmos, Richard Hill, Ben Lotto, and Wade Ramey.

texts offered for sale, only one obscure book gave a determinant-free proof that eigenvalues exist, and that book did not manage to develop other key parts of linear algebra without determinants. The anti-determinant philosophy advocated in this paper is an attempt to counter the undeserved dominance of determinant-dependent methods.

This paper focuses on showing that determinants should be banished from much of the theoretical part of linear algebra. Determinants are also useless in the computational part of linear algebra. For example, Cramer's rule for solving systems of linear equations is already worthless for 10×10 systems, not to mention the much larger systems often encountered in the real world. Many computer programs efficiently calculate eigenvalues numerically—none of them uses determinants. To emphasize the point, let me quote a numerical analyst. Henry Thacher, in a review (*SIAM News*, September 1988) of the *Turbo Pascal Numerical Methods Toolbox*, writes,

I find it hard to conceive of a situation in which the numerical value of a determinant is needed: Cramer's rule, because of its inefficiency, is completely impractical, while the magnitude of the determinant is an indication of neither the condition of the matrix nor the accuracy of the solution.

2. EIGENVALUES AND EIGENVECTORS. The basic objects of study in linear algebra can be thought of as either linear transformations or matrices. Because a basis-free approach seems more natural, this paper will mostly use the language of linear transformations; readers who prefer the language of matrices should have no trouble making the appropriate translation. The term *linear operator* will mean a linear transformation from a vector space to itself; thus a linear operator corresponds to a square matrix (assuming some choice of basis).

Notation used throughout the paper: n denotes a positive integer, V denotes an n -dimensional complex vector space, T denotes a linear operator on V , and I denotes the identity operator.

A complex number λ is called an *eigenvalue* of T if $T - \lambda I$ is not injective. Here is the central result about eigenvalues, with a simple proof that avoids determinants.

Theorem 2.1. *Every linear operator on a finite-dimensional complex vector space has an eigenvalue.*

Proof: To show that T (our linear operator on V) has an eigenvalue, fix any non-zero vector $v \in V$. The vectors $v, Tv, T^2v, \dots, T^n v$ cannot be linearly independent, because V has dimension n and we have $n + 1$ vectors. Thus there exist complex numbers a_0, \dots, a_n , not all 0, such that

$$a_0 v + a_1 Tv + \dots + a_n T^n v = 0.$$

Make the a 's the coefficients of a polynomial, which can be written in factored form as

$$a_0 + a_1 z + \dots + a_n z^n = c(z - r_1) \cdots (z - r_m),$$

where c is a non-zero complex number, each r_j is complex, and the equation holds for all complex z . We then have

$$\begin{aligned} 0 &= (a_0 I + a_1 T + \dots + a_n T^n) v \\ &= c(T - r_1 I) \cdots (T - r_m I) v, \end{aligned}$$

which means that $T - r_j I$ is not injective for at least one j . In other words, T has an eigenvalue. \square

Recall that a vector $v \in V$ is called an *eigenvector* of T if $Tv = \lambda v$ for some eigenvalue λ . The next proposition—which has a simple, determinant-free proof—obviously implies that the number of distinct eigenvalues of T cannot exceed the dimension of V .

Proposition 2.2. *Non-zero eigenvectors corresponding to distinct eigenvalues of T are linearly independent.*

Proof: Suppose that v_1, \dots, v_m are non-zero eigenvectors of T corresponding to distinct eigenvalues $\lambda_1, \dots, \lambda_m$. We need to prove that v_1, \dots, v_m are linearly independent. To do this, suppose a_1, \dots, a_m are complex numbers such that

$$a_1 v_1 + \dots + a_m v_m = 0.$$

Apply the linear operator $(T - \lambda_2 I)(T - \lambda_3 I) \cdots (T - \lambda_m I)$ to both sides of the equation above, getting

$$a_1(\lambda_1 - \lambda_2)(\lambda_1 - \lambda_3) \cdots (\lambda_1 - \lambda_m)v_1 = 0.$$

Thus $a_1 = 0$. In a similar fashion, $a_j = 0$ for each j , as desired. \square

3. GENERALIZED EIGENVECTORS. Unfortunately, the eigenvectors of T need not span V . For example, the linear operator on \mathbb{C}^2 whose matrix is

$$\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

has only one eigenvalue, namely 0, and its eigenvectors form a one-dimensional subspace of \mathbb{C}^2 . We will see, however, that the generalized eigenvectors (defined below) of T always span V .

A vector $v \in V$ is called a *generalized eigenvector* of T if

$$(T - \lambda I)^k v = 0$$

for some eigenvalue λ of T and some positive integer k . Obviously, the set of generalized eigenvectors of T corresponding to an eigenvalue λ is a subspace of V . The following lemma shows that in the definition of generalized eigenvector, instead of allowing an arbitrary power of $T - \lambda I$ to annihilate v , we could have restricted attention to the n^{th} power, where n equals the dimension of V . As usual, \ker is an abbreviation for kernel (the set of vectors that get mapped 0).

Lemma 3.1. *The set of generalized eigenvectors of T corresponding to an eigenvalue λ equals $\ker(T - \lambda I)^n$.*

Proof: Obviously, every element of $\ker(T - \lambda I)^n$ is a generalized eigenvector of T corresponding to λ . To prove the inclusion in the other direction, let v be a generalized eigenvector of T corresponding to λ . We need to prove that $(T - \lambda I)^n v = 0$. Clearly, we can assume that $v \neq 0$, so there is a smallest non-negative integer k such that $(T - \lambda I)^k v = 0$. We will be done if we show that $k \leq n$. This will be proved by showing that

$$v, (T - \lambda I)v, (T - \lambda I)^2 v, \dots, (T - \lambda I)^{k-1} v \quad (3.2)$$

are linearly independent vectors; we will then have k linearly independent elements in an n -dimensional space, which implies that $k \leq n$.

To prove the vectors in (3.2) are linearly independent, suppose a_0, \dots, a_{k-1} are complex numbers such that

$$a_0 v + a_1(T - \lambda I)v + \cdots + a_{k-1}(T - \lambda I)^{k-1}v = 0. \quad (3.3)$$

Apply $(T - \lambda I)^{k-1}$ to both sides of the equation above, getting $a_0(T - \lambda I)^{k-1}v = 0$, which implies that $a_0 = 0$. Now apply $(T - \lambda I)^{k-2}$ to both sides of (3.3), getting $a_1(T - \lambda I)^{k-1}v = 0$, which implies that $a_1 = 0$. Continuing in this fashion, we see that $a_j = 0$ for each j , as desired. \square

The next result is the key tool we'll use to give a description of the structure of a linear operator.

Proposition 3.4. *The generalized eigenvectors of T span V .*

Proof: The proof will be by induction on n , the dimension of V . Obviously, the result holds when $n = 1$.

Suppose that $n > 1$ and that the result holds for all vector spaces of dimension less than n . Let λ be any eigenvalue of T (one exists by Theorem 2.1). We first show that

$$V = \underbrace{\ker(T - \lambda I)^n}_{V_1} \oplus \underbrace{\text{ran}(T - \lambda I)^n}_{V_2}; \quad (3.5)$$

here, as usual, ran is an abbreviation for range. To prove (3.5), suppose $v \in V_1 \cap V_2$. Then $(T - \lambda I)^n v = 0$ and there exists $u \in V$ such that $(T - \lambda I)^n u = v$. Applying $(T - \lambda I)^n$ to both sides of the last equation, we have $(T - \lambda I)^{2n} u = 0$. This implies that $(T - \lambda I)^n u = 0$ (by Lemma 3.1), which implies that $v = 0$. Thus

$$V_1 \cap V_2 = \{0\}. \quad (3.6)$$

Because V_1 and V_2 are the kernel and range of a linear operator on V , we have

$$\dim V = \dim V_1 + \dim V_2. \quad (3.7)$$

Equations (3.6) and (3.7) imply (3.5).

Note that $V_1 \neq \{0\}$ (because λ is an eigenvalue of T), and thus $\dim V_2 < n$. Furthermore, because T commutes with $(T - \lambda I)^n$, we easily see that T maps V_2 into V_2 . By our induction hypothesis, V_2 is spanned by the generalized eigenvectors of $T|_{V_2}$, each of which is obviously also a generalized eigenvector of T . Everything in V_1 is a generalized eigenvector of T , and hence (3.5) gives the desired result. \square

A nice corollary of the last proposition is that if 0 is the only eigenvalue of T , then T is nilpotent (recall that an operator is called *nilpotent* if some power of it equals 0). Proof: If 0 is the only eigenvalue of T , then every vector in V is a generalized eigenvector of T corresponding to the eigenvalue 0 (by Proposition 3.4); Lemma 3.1 then implies that $T^n = 0$.

Non-zero eigenvectors corresponding to distinct eigenvalues are linearly independent (Proposition 2.2). We need an analogous result with generalized eigenvectors replacing eigenvectors. This can be proved by following the basic pattern of the proof of Proposition 2.2, as we now do.

Proposition 3.8. *Non-zero generalized eigenvectors corresponding to distinct eigenvalues of T are linearly independent.*

Proof: Suppose that v_1, \dots, v_m are non-zero generalized eigenvectors of T corresponding to distinct eigenvalues $\lambda_1, \dots, \lambda_m$. We need to prove that v_1, \dots, v_m are linearly independent. To do this, suppose a_1, \dots, a_m are complex numbers such that

$$a_1 v_1 + \dots + a_m v_m = 0. \quad (3.9)$$

Let k be the smallest positive integer such that $(T - \lambda_1 I)^k v_1 = 0$. Apply the linear operator

$$(T - \lambda_1 I)^{k-1} (T - \lambda_2 I)^n \dots (T - \lambda_m I)^n$$

to both sides of (3.9), getting

$$a_1 (T - \lambda_1 I)^{k-1} (T - \lambda_2 I)^n \dots (T - \lambda_m I)^n v_1 = 0, \quad (3.10)$$

where we have used Lemma 3.1. If we rewrite $(T - \lambda_2 I)^n \dots (T - \lambda_m I)^n$ in (3.10) as

$$((T - \lambda_1 I) + (\lambda_1 - \lambda_2)I)^n \dots ((T - \lambda_1 I) + (\lambda_1 - \lambda_m)I)^n,$$

and then expand each $((T - \lambda_1 I) + (\lambda_1 - \lambda_j)I)^n$ using the binomial theorem and multiply everything together, we get a sum of terms. Except for the term

$$(\lambda_1 - \lambda_2)^n \dots (\lambda_1 - \lambda_m)^n I,$$

each term in this sum includes a power of $(T - \lambda_1 I)$, which when combined with the $(T - \lambda_1 I)^{k-1}$ on the left and the v_1 on the right in (3.10) gives 0. Hence (3.10) becomes the equation

$$a_1 (\lambda_1 - \lambda_2)^n \dots (\lambda_1 - \lambda_m)^n (T - \lambda_1 I)^{k-1} v_1 = 0.$$

Thus $a_1 = 0$. In a similar fashion, $a_j = 0$ for each j , as desired. \square

Now we can pull everything together into the following structure theorem. Part (b) allows us to interpret each linear transformation appearing in parts (c) and (d) as a linear operator from U_j to itself.

Theorem 3.11. *Let $\lambda_1, \dots, \lambda_m$ be the distinct eigenvalues of T , with U_1, \dots, U_m denoting the corresponding sets of generalized eigenvectors. Then*

- (a) $V = U_1 \oplus \dots \oplus U_m$;
- (b) T maps each U_j into itself;
- (c) each $(T - \lambda_j I)|_{U_j}$ is nilpotent;
- (d) each $T|_{U_j}$ has only one eigenvalue, namely λ_j .

Proof: The proof of (a) follows immediately from Propositions 3.8 and 3.4.

To prove (b), suppose $v \in U_j$. Then $(T - \lambda_j I)^k v = 0$ for some positive integer k . We have

$$(T - \lambda_j I)^k T v = T(T - \lambda_j I)^k v = T(0) = 0.$$

Thus $T v \in U_j$, as desired.

The proof of (c) follows immediately from the definition of a generalized eigenvector and Lemma 3.1.

To prove (d), let λ be an eigenvalue of $T|_{U_j}$ with corresponding non-zero eigenvector $v \in U_j$. Then $(T - \lambda_j I)v = (\lambda - \lambda_j)v$, and hence

$$(T - \lambda_j I)^k v = (\lambda - \lambda_j)^k v$$

for each positive integer k . Because v is a generalized eigenvector of T corresponding to λ_j , the left-hand side of this equation is 0 for some k . Thus $\lambda = \lambda_j$, as desired. \square

4. THE MINIMAL POLYNOMIAL. Because the space of linear operators on V is finite dimensional, there is a smallest positive integer k such that

$$I, T, T^2, \dots, T^k$$

are not linearly independent. Thus there exist unique complex numbers a_0, \dots, a_{k-1} such that

$$a_0 I + a_1 T + a_2 T^2 + \dots + a_{k-1} T^{k-1} + T^k = 0.$$

The polynomial

$$a_0 + a_1 z + a_2 z^2 + \dots + a_{k-1} z^{k-1} + z^k$$

is called the *minimal polynomial* of T . It is the monic polynomial p of smallest degree such that $p(T) = 0$ (a *monic polynomial* is one whose term of highest degree has coefficient 1).

The next theorem connects the minimal polynomial to the decomposition of V as a direct sum of generalized eigenvectors.

Theorem 4.1. *Let $\lambda_1, \dots, \lambda_m$ be the distinct eigenvalues of T , let U_j denote the set of generalized eigenvectors corresponding to λ_j , and let α_j be the smallest positive integer such that $(T - \lambda_j I)^{\alpha_j} v = 0$ for every $v \in U_j$. Let*

$$p(z) = (z - \lambda_1)^{\alpha_1} \dots (z - \lambda_m)^{\alpha_m}. \quad (4.2)$$

Then

- (a) p is the minimal polynomial of T ;
- (b) p has degree at most $\dim V$;
- (c) if q is a polynomial such that $q(T) = 0$, then q is a polynomial multiple of p .

Proof: We will prove first (b), then (c), then (a).

To prove (b), note that each α_j is at most the dimension of U_j (by Lemma 3.1 applied to $T|_{U_j}$). Because $V = U_1 \oplus \dots \oplus U_m$ (by Theorem 3.11(a)), the α_j 's can add up to at most the dimension of V . Thus (b) holds.

To prove (c), suppose q is a polynomial such that $q(T) = 0$. If we show that q is a polynomial multiple of each $(z - \lambda_j)^{\alpha_j}$, then (c) will hold. To do this, fix j . The polynomial q has the form

$$q(z) = c(z - r_1)^{\delta_1} \dots (z - r_M)^{\delta_M} (z - \lambda_j)^{\delta},$$

where $c \in \mathbb{C}$, the r_k 's are complex numbers all different from λ_j , the δ_k 's are positive integers, and δ is a non-negative integer. If $c = 0$, we are done, so assume that $c \neq 0$. Suppose $v \in U_j$. Then $(T - \lambda_j I)^{\delta} v$ is also in U_j (by Theorem 3.11(b)). Now

$$c(T - r_1 I)^{\delta_1} \dots (T - r_M I)^{\delta_M} (T - \lambda_j I)^{\delta} v = q(T)v = 0,$$

and $(T - r_1 I)^{\delta_1} \cdots (T - r_M I)^{\delta_M}$ is injective on U_j (by Theorem 3.11(d)). Thus $(T - \lambda_j I)^{\delta_j} v = 0$. Because v was an arbitrary element of U_j , this implies that $\alpha_j \leq \delta_j$. Thus q is a polynomial multiple of $(z - \lambda_j)^{\alpha_j}$, and (c) holds.

To prove (a), suppose v is a vector in some U_j . If we commute the terms of $(T - \lambda_1 I)^{\alpha_1} \cdots (T - \lambda_m I)^{\alpha_m}$ (which equals $p(T)$) so that $(T - \lambda_j I)^{\alpha_j}$ is on the right, we see that $p(T)v = 0$. Because U_1, \dots, U_m span V (Theorem 3.11(a)), we conclude that $p(T) = 0$. In other words, p is a monic polynomial that annihilates T . We know from (c) that no monic polynomial of lower degree has this property. Thus p must be the minimal polynomial of T , completing the proof. \square

Note that by avoiding determinants we have been naturally led to the description of the minimal polynomial in terms of generalized eigenvectors.

5. MULTIPLICITY AND THE CHARACTERISTIC POLYNOMIAL. The *multiplicity* of an eigenvalue λ of T is defined to be the dimension of the set of generalized eigenvectors of T corresponding to λ . We see immediately that the sum of the multiplicities of all eigenvalues of T equals n , the dimension of V (from Theorem 3.11(a)). Note that the definition of multiplicity given here has a clear connection with the geometric behavior of T , whereas the usual definition (as the multiplicity of a root of the polynomial $\det(zI - T)$) describes an object without obvious meaning.

Let $\lambda_1, \dots, \lambda_m$ denote the distinct eigenvalues of T , with corresponding multiplicities β_1, \dots, β_m . The polynomial

$$(z - \lambda_1)^{\beta_1} \cdots (z - \lambda_m)^{\beta_m} \quad (5.1)$$

is called the *characteristic polynomial* of T . Clearly, it is a polynomial of degree n .

Of course the usual definition of the characteristic polynomial involves a determinant; the characteristic polynomial is then used to prove the existence of eigenvalues. Without mentioning determinants, we have reversed that procedure. We first showed that T has n eigenvalues, counting multiplicities, and then used that to give a more natural definition of the characteristic polynomial (“counting multiplicities” means that each eigenvalue is repeated as many times as its multiplicity).

The next result is called the Cayley-Hamilton Theorem. With the approach taken here, its proof is easy.

Theorem 5.2. *Let q denote the characteristic polynomial of T . Then $q(T) = 0$.*

Proof: Let U_j and α_j be as in Theorem 4.1, and let β_j equal the dimension of U_j . As we noted earlier, $\alpha_j \leq \beta_j$ (by Lemma 3.1 applied to $T|_{U_j}$). Hence the characteristic polynomial (5.1) is a polynomial multiple of the minimal polynomial (4.2). Thus the characteristic polynomial must annihilate T . \square

6. UPPER-TRIANGULAR FORM. A square matrix is called *upper-triangular* if all the entries below the main diagonal are 0. Our next goal is to show that each linear operator has an upper-triangular matrix for some choice of basis. We’ll begin with nilpotent operators; our main structure theorem will then easily give the result for arbitrary linear operators.

Lemma 6.1. Suppose T is nilpotent. Then there is a basis of V with respect to which the matrix of T contains only 0's on and below the main diagonal.

Proof: First choose a basis of $\ker T$. Then extend this to a basis of $\ker T^2$. Then extend to a basis of $\ker T^3$. Continue in this fashion, eventually getting a basis of V . The matrix of T with respect to this basis clearly has the desired form. \square

By avoiding determinants and focusing on generalized eigenvectors, we can now give a simple proof that every linear operator can be put in upper-triangular form. We actually get a better result, because the matrix in the next theorem has many more 0's than required for upper-triangular form.

Theorem 6.2. Let $\lambda_1, \dots, \lambda_m$ be the distinct eigenvalues of T . Then there is a basis of V with respect to which the matrix of T has the form

$$\begin{bmatrix} \boxed{\begin{matrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_1 \end{matrix}} & & & \\ & \boxed{\begin{matrix} \lambda_2 & & * \\ & \ddots & \\ 0 & & \lambda_2 \end{matrix}} & & \\ & & \ddots & \\ & & & \boxed{\begin{matrix} \lambda_m & & * \\ & \ddots & \\ 0 & & \lambda_m \end{matrix}} \end{bmatrix}.$$

Proof: This follows immediately from Theorem 3.11 and Lemma 6.1. \square

For many traditional uses of the Jordan form, the theorem above can be used instead. If Jordan form really is needed, then many standard proofs show (without determinants) that every nilpotent operator can be put in Jordan form. The result for general linear operators then follows from Theorem 3.11.

7. THE SPECTRAL THEOREM. In this section we assume that \langle, \rangle is an inner product on V . The nicest linear operators on V are those for which there is an orthonormal basis of V consisting of eigenvectors. With respect to any such basis, the matrix of the linear operator is diagonal, meaning that it is 0 everywhere except along the main diagonal, which must contain the eigenvalues. The Spectral Theorem, which we'll prove in this section, describes precisely those linear operators for which there is an orthonormal basis of V consisting of eigenvectors.

Recall that the *adjoint* of T is the unique linear operator T^* on V such that

$$\langle Tu, v \rangle = \langle u, T^*v \rangle$$

for all $u, v \in V$. The linear operator T is called *normal* if T commutes with its adjoint; in other words, T is normal if $TT^* = T^*T$. The linear operator T is called *self-adjoint* if $T = T^*$. Obviously, every self-adjoint operator is normal. We'll see

that the normal operators are precisely the ones that can be diagonalized by an orthonormal basis. Before proving that, we need a few preliminary results. Note that the next lemma is trivial if T is self-adjoint.

Lemma 7.1. *If T is normal, then $\ker T = \ker T^*$.*

Proof: If T is normal and $v \in V$, then

$$\langle Tv, Tv \rangle = \langle T^*Tv, v \rangle = \langle TT^*v, v \rangle = \langle T^*v, T^*v \rangle.$$

Thus $Tv = 0$ if and only if $T^*v = 0$. □

The next proposition, combined with our result that the generalized eigenvectors of a linear operator span the domain (Proposition 3.4), shows that the eigenvectors of a normal operator span the domain.

Proposition 7.2. *Every generalized eigenvector of a normal operator is an eigenvector of the operator.*

Proof: Suppose T is normal. We will prove that

$$\ker T^k = \ker T \tag{7.3}$$

for every positive integer k . This will complete the proof of the proposition, because we can replace T in (7.3) by $T - \lambda I$ for arbitrary $\lambda \in \mathbb{C}$.

We prove (7.3) by induction on k . Clearly, the result holds for $k = 1$. Suppose now that k is a positive integer such that (7.3) holds. Let $v \in \ker T^{k+1}$. Then $T(T^k v) = T^{k+1}v = 0$. In other words, $T^k v \in \ker T$, and so $T^*(T^k v) = 0$ (by Lemma 7.1). Thus

$$0 = \langle T^*(T^k v), T^{k-1}v \rangle = \langle T^k v, T^k v \rangle.$$

Hence $v \in \ker T^k$, which implies that $v \in \ker T$ (by our induction hypothesis). Thus $\ker T^{k+1} = \ker T$, completing the induction. □

The last proposition, together with Proposition 3.4, implies that a normal operator can be diagonalized by some basis. The next proposition will be used to show that this can be done by an orthonormal basis.

Proposition 7.4. *Eigenvectors of a normal operator corresponding to distinct eigenvalues are orthogonal.*

Proof: Suppose T is normal and α, λ are distinct eigenvalues of T , with corresponding eigenvectors u, v . Thus $(T - \lambda I)v = 0$, and so $(T^* - \bar{\lambda}I)v = 0$ (by Lemma 7.1). In other words, v is also an eigenvector of T^* , with eigenvalue $\bar{\lambda}$. Now

$$\begin{aligned} (\alpha - \lambda)\langle u, v \rangle &= \langle \alpha u, v \rangle - \langle u, \bar{\lambda}v \rangle \\ &= \langle Tu, v \rangle - \langle u, T^*v \rangle \\ &= \langle Tu, v \rangle - \langle Tu, v \rangle \\ &= 0. \end{aligned}$$

Thus $\langle u, v \rangle = 0$, as desired. □

Now we can put everything together, getting the finite-dimensional Spectral Theorem for complex inner product spaces.

Theorem 7.5. *There is an orthonormal basis of V consisting of eigenvectors of T if and only if T is normal.*

Proof: To prove the easy direction, first suppose that there is an orthonormal basis of V consisting of eigenvectors of T . With respect to that basis, T has a diagonal matrix. The matrix of T^* (with respect to the same basis) is obtained by taking the conjugate transpose of the matrix of T ; hence T^* also has a diagonal matrix. Any two diagonal matrices commute. Thus T commutes with T^* , which means that T is normal.

To prove the other direction, now suppose that T is normal. For each eigenvalue of T , choose an orthonormal basis of the associated set of eigenvectors. The union of these bases (one for each eigenvalue) is still an orthonormal set, because eigenvectors corresponding to distinct eigenvalues are orthogonal (by Proposition 7.4). The span of this union includes every eigenvector of T (by construction), and hence every generalized eigenvector of T (by Proposition 7.2). But the generalized eigenvectors of T span V (by Proposition 3.4), and so we have an orthonormal basis of V consisting of eigenvectors of T . \square

The proposition below will be needed in the next section, when we prove the Spectral Theorem for real inner product spaces.

Proposition 7.6. *Every eigenvalue of a self-adjoint operator is real.*

Proof: Suppose T is self-adjoint. Let λ be an eigenvalue of T , and let v be a non-zero vector in V such that $Tv = \lambda v$. Then

$$\lambda\|v\|^2 = \langle \lambda v, v \rangle = \langle Tv, v \rangle = \langle v, Tv \rangle = \langle v, \lambda v \rangle = \bar{\lambda}\|v\|^2.$$

Thus $\lambda = \bar{\lambda}$, which means that λ is real, as desired. \square

8. GETTING REAL. So far we have been dealing only with complex vector spaces. As we'll see, a real vector space U can be embedded, in a natural way, in a complex vector space called the complexification of U . Each linear operator on U can be extended to a linear operator on the complexification of U . Our results about linear operators on complex vector spaces can then be translated to information about linear operators on real vector spaces. Let's see how this process works.

Suppose that U is a real vector space. As a set, the *complexification* of U , denoted $U_{\mathbb{C}}$, equals $U \times U$. Formally, a typical element of $U_{\mathbb{C}}$ is an ordered pair (u, v) , where $u, v \in U$, but we will write this as $u + iv$, for obvious reasons. We define addition on $U_{\mathbb{C}}$ by

$$(u_1 + iv_1) + (u_2 + iv_2) = (u_1 + u_2) + i(v_1 + v_2).$$

The notation shows how we should define multiplication by complex scalar on $U_{\mathbb{C}}$:

$$(a + ib)(u + iv) = (au - bv) + i(av + bu)$$

for $a, b \in \mathbb{R}$ and $u, v \in U$. With these definitions of addition and scalar multiplication, $U_{\mathbb{C}}$ becomes a complex vector space. We can think of U as a subset of $U_{\mathbb{C}}$ by identifying $u \in U$ with $u + i0$. Clearly, any basis of the real vector space U is

also a basis of the complex vector space $U_{\mathbb{C}}$. Hence the dimension of U as a real vector space equals the dimension of $U_{\mathbb{C}}$ as a complex vector space.

For S a linear operator on a real vector space U , the complexification of S , denoted $S_{\mathbb{C}}$, is the linear operator on $U_{\mathbb{C}}$ defined by

$$S_{\mathbb{C}}(u + iv) = Su + iSv$$

for $u, v \in U$. If we choose a basis of U and also think of it as a basis of $U_{\mathbb{C}}$, then clearly S and $S_{\mathbb{C}}$ have the same matrix with respect to this basis.

Note that any real eigenvalue of $S_{\mathbb{C}}$ is also an eigenvalue of S (because if $a \in \mathbb{R}$ and $S_{\mathbb{C}}(u + iv) = a(u + iv)$, then $Su = au$ and $Sv = av$). Non-real eigenvalues of $S_{\mathbb{C}}$ come in pairs. More precisely,

$$(S_{\mathbb{C}} - \lambda I)^j(u + iv) = 0 \Leftrightarrow (S_{\mathbb{C}} - \bar{\lambda} I)^j(u - iv) = 0 \quad (8.1)$$

for j a positive integer, $\lambda \in \mathbb{C}$, and $u, v \in U$, as is easily proved by induction on j . In particular, if $\lambda \in \mathbb{C}$ is an eigenvalue of $S_{\mathbb{C}}$, then so is $\bar{\lambda}$, and the multiplicity of λ (recall that this is defined as the dimension of the set of generalized eigenvectors of $S_{\mathbb{C}}$ corresponding to λ) is the same as the multiplicity of $\bar{\lambda}$. Because the sum of the multiplicities of all the eigenvalues of $S_{\mathbb{C}}$ equals the (complex) dimension of $U_{\mathbb{C}}$ (by Theorem 3.11(a)), we see that if $U_{\mathbb{C}}$ has odd (complex) dimension, then $S_{\mathbb{C}}$ must have a real eigenvalue. Putting all this together, we have proved the following theorem. Once again, a proof without determinants offers more insight into why the result holds than the standard proof using determinants.

Theorem 8.2. *Every linear operator on an odd-dimensional real vector space has a real eigenvalue.*

The minimal and characteristic polynomials of a linear operator S on a real vector space are defined to be the corresponding polynomials of the complexification $S_{\mathbb{C}}$. Both these polynomials have real coefficients—this follows from our definitions of minimal and characteristic polynomials and (8.1). The reader should be able to derive the properties of these polynomials easily from the corresponding results on complex vector spaces (Theorems 4.1 and 5.2).

Our procedure for transferring results from complex vector spaces to real vector spaces can also be used to prove the real Spectral Theorem. To see how that works, suppose now that U is a real inner product space with inner product $\langle \cdot, \cdot \rangle$. We make the complexification $U_{\mathbb{C}}$ into a complex inner product space by defining an inner product on $U_{\mathbb{C}}$ in the obvious way:

$$\langle u_1 + iv_1, u_2 + iv_2 \rangle = \langle u_1, u_2 \rangle + \langle v_1, v_2 \rangle + i\langle v_1, u_2 \rangle - i\langle u_1, v_2 \rangle.$$

Note that any orthonormal basis of the real inner product space U is also an orthonormal basis of the complex inner product space $U_{\mathbb{C}}$.

If S is a self-adjoint operator on U , then obviously $S_{\mathbb{C}}$ is self-adjoint on $U_{\mathbb{C}}$. We can then apply the complex Spectral Theorem (Theorem 7.5) to $S_{\mathbb{C}}$ and transfer to U , getting the real Spectral Theorem. The next theorem gives the formal statement of the result and the details of the proof.

Theorem 8.3. *Suppose U is a real inner product space and S is a linear operator on U . Then there is an orthonormal basis of U consisting of eigenvectors of S if and only if S is self-adjoint.*

Proof: To prove the easy direction, first suppose that there is an orthonormal basis of U consisting of eigenvectors of S . With respect to that basis, S has a diagonal matrix. Clearly, the matrix of S^* (with respect to the same basis) equals the matrix of S . Thus S is self-adjoint.

To prove the other direction, now suppose that S is self-adjoint. As noted above, this implies that S_C is self-adjoint on U_C . Thus there is a basis

$$\{u_1 + iv_1, \dots, u_n + iv_n\} \quad (8.4)$$

of U_C consisting of eigenvectors of S_C (by the complex Spectral Theorem, which is Theorem 7.5); here each u_j and v_j is in U . Each eigenvalue of S_C is real (Proposition 7.6), and thus each u_j and each v_j is an eigenvector of S . Clearly, $\{u_1, v_1, \dots, u_n, v_n\}$ spans U (because (8.4) is a basis of U_C). Conclusion: The eigenvectors of S span U .

For each eigenvalue of S , choose an orthonormal basis of the associated set of eigenvectors in U . The union of these bases (one for each eigenvalue) is still orthonormal, because eigenvectors corresponding to distinct eigenvalues are orthogonal (Proposition 7.4). The span of this union includes every eigenvector of S (by construction). We have just seen that the eigenvectors of S span U , and so we have an orthonormal basis of U consisting of eigenvectors of S , as desired. \square

9. DETERMINANTS. At this stage we have proved most of the major structure theorems of linear algebra without even defining determinants. In this section we will give a simple definition of determinants, whose main reasonable use in undergraduate mathematics is in the change of variables formula for multi-variable integrals.

The constant term of the characteristic polynomial of T is plus or minus the product of the eigenvalues of T , counting multiplicity (this is obvious from our definition of the characteristic polynomial). Let's look at some additional motivation for studying the product of the eigenvalues.

Suppose we want to know how to make a change of variables in a multi-variable integral over some subset of \mathbf{R}^n . After linearization, this reduces to the question of how a linear operator S on \mathbf{R}^n changes volumes. Let's consider the special case where S is self-adjoint. Then there is an orthonormal basis of \mathbf{R}^n consisting of eigenvectors of S (by the real Spectral Theorem, which is Theorem 8.3). A moment's thought about the geometry of an orthonormal basis of eigenvectors shows that if E is a subset of \mathbf{R}^n , then the volume (whatever that means) of $S(E)$ must equal the volume of E multiplied by the absolute value of the product of the eigenvalues of S , counting multiplicity. We'll prove later that a similar result holds even for non-self-adjoint operators. At any rate, we see that the product of the eigenvalues seems to be an interesting object. An arbitrary linear operator on a real vector space need not have any eigenvalues, so we will return to our familiar setting of a linear operator T on a complex vector space V . After getting the basic results on complex vector spaces, we'll deal with real vector spaces by using the notion of complexification discussed earlier.

Now we are ready for the formal definition. The *determinant* of T , denoted $\det T$, is defined to be the product of the eigenvalues of T , counting multiplicity. This definition would not be possible with the traditional approach to eigenvalues, because that method uses determinants to prove that eigenvalues exist. With the techniques used here, we already know (by Theorem 3.11(a)) that T has $\dim V$ eigenvalues, counting multiplicity. Thus our simple definition makes sense.

In addition to simplicity, our definition also makes transparent the following result, which is not at all obvious from the standard definition.

Theorem 9.1. *An operator is invertible if and only if its determinant is non-zero.*

Proof: Clearly, T is invertible if and only if 0 is not an eigenvalue of T , and this happens if and only if $\det T \neq 0$. \square

With our definition of determinant and characteristic polynomial, we see immediately that the constant term of the characteristic polynomial of T equals $(-1)^n \det T$, where $n = \dim V$. The next result shows that even more is true—our definitions are consistent with the usual ones.

Proposition 9.2. *The characteristic polynomial of T equals $\det(zI - T)$.*

Proof: Let $\lambda_1, \dots, \lambda_m$ denote the eigenvalues of T , with multiplicities β_1, \dots, β_m . Thus for $z \in \mathbb{C}$, the eigenvalues of $zI - T$ are $z - \lambda_1, \dots, z - \lambda_m$, with multiplicities β_1, \dots, β_m . Hence the determinant of $zI - T$ is the product

$$(z - \lambda_1)^{\beta_1} \cdots (z - \lambda_m)^{\beta_m},$$

which equals the characteristic polynomial of T . \square

Note that determinant is a similarity invariant. In other words, if S is an invertible linear operator on V , then T and STS^{-1} have the same determinant (because they have the same eigenvalues, counting multiplicity).

We define the determinant of a square matrix of complex numbers to be the determinant of the corresponding linear operator (with respect to some choice of basis, which doesn't matter, because two different bases give rise to two linear operators that are similar and hence have the same determinant). Fix a basis of V , and for the rest of this section let's identify linear operators on V with matrices with respect to that basis. How can we find the determinant of T from its matrix, without finding all the eigenvalues? Although getting the answer to that question will be hard, the method used below will show how someone might have discovered the formula for the determinant of a matrix. Even with the derivation that follows, determinants are difficult, which is precisely why they should be avoided.

We begin our search for a formula for the determinant by considering matrices of a special form. Let $a_1, \dots, a_n \in \mathbb{C}$. Consider a linear operator T whose matrix is

$$\begin{bmatrix} 0 & & & & a_n \\ a_1 & 0 & & & \\ & a_2 & 0 & & \\ & & \ddots & \ddots & \\ & & & a_{n-1} & 0 \end{bmatrix}; \quad (9.3)$$

here all entries of the matrix are 0 except for the upper right-hand corner and along the line just below the main diagonal. Let's find the determinant of T . Note that $T^n = a_1 \cdots a_n I$. Because the first columns of $\{I, T, \dots, T^{n-1}\}$ are linearly independent (assuming that none of the a_j is 0), no polynomial of degree less than n can annihilate T . Thus $z^n - a_1 \cdots a_n$ is the minimal polynomial of T . Hence

$z^n - a_1 \cdots a_n$ is also the characteristic polynomial of T . Thus

$$\det T = (-1)^{n-1} a_1 \cdots a_n.$$

(If some a_j is 0, then clearly T is not invertible, so $\det T = 0$, and the same formula holds.)

Now let τ be a permutation of $\{1, \dots, n\}$, and consider a matrix T whose j^{th} column consists of all zeroes except for a_j in the $\tau(j)^{\text{th}}$ row. The permutation τ is a product of cyclic permutations. Thus T is similar to (and so has the same determinant as) a block diagonal matrix where each block of size greater than one has the form of (9.3). The determinant of a block diagonal matrix is obviously the product of the determinants of the blocks, and we know from the last paragraph how to compute those. Thus we see that $\det T = (\text{sign } \tau) a_1 \cdots a_n$. To put this into a form that does not depend upon the particular permutation τ , let $t_{i,j}$ denote the entry in row i , column j , of T (so $t_{i,j} = 0$ unless $i = \tau(j)$), and let $P(n)$ denote the set of all permutations of $\{1, \dots, n\}$. Then

$$\det T = \sum_{\pi \in P(n)} (\text{sign } \pi) t_{\pi(1),1} \cdots t_{\pi(n),n}, \quad (9.4)$$

because each summand is 0 except the one corresponding to the permutation τ .

Consider now an arbitrary matrix T with entries $t_{i,j}$. Using the paragraph above as motivation, we guess that the formula for $\det T$ is given by (9.4). The next proposition shows that this guess is correct and gives the usual formula for the determinant of a matrix.

Proposition 9.5. $\det(T) = \sum_{\pi \in P(n)} (\text{sign } \pi) t_{\pi(1),1} \cdots t_{\pi(n),n}$.

Proof: Define a function d on the set of $n \times n$ matrices by

$$d(T) = \sum_{\pi \in P(n)} (\text{sign } \pi) t_{\pi(1),1} \cdots t_{\pi(n),n}.$$

We want to prove that $\det T = d(T)$. To do this, choose S so that STS^{-1} is in the upper triangular form given by Theorem 6.2. Now $d(STS^{-1})$ equals the product of the entries on the main diagonal of STS^{-1} (because only the identity permutation makes a non-zero contribution to the sum defining $d(STS^{-1})$). But the entries on the main diagonal of STS^{-1} are precisely the eigenvalues of T , counting multiplicity, so $\det T = d(STS^{-1})$. Thus to complete the proof, we need only show that d is a similarity invariant; then we will have $\det T = d(STS^{-1}) = d(T)$.

To show that d is a similarity invariant, first prove that d is multiplicative, meaning that $d(AB) = d(A)d(B)$ for all $n \times n$ matrices A and B . The proof that d is multiplicative, which will not be given here, consists of a straightforward rearrangement of terms appearing in the formula defining $d(AB)$ (see any text that defines $\det(T)$ to be $d(T)$ and then proves that $\det AB = (\det A)(\det B)$). The multiplicativity of d now leads to a proof that d is a similarity invariant, as follows:

$$d(STS^{-1}) = d(ST)d(S^{-1}) = d(S^{-1})d(ST) = d(S^{-1}ST) = d(T).$$

Thus $\det T = d(T)$, as claimed. \square

All the usual properties of determinants can be proved either from the (new) definition or from Proposition 9.5. In particular, the last proof shows that \det is multiplicative.

The determinant of a linear operator on a real vector space is defined to be the determinant (product of the eigenvalues) of its complexification. Proposition 9.5 holds on real as well as complex vector spaces. To see this, suppose that U is a real vector space and S is a linear operator on U . If we choose a basis of U and also think of it as a basis of the complexification $U_{\mathbb{C}}$, then S and its complexification $S_{\mathbb{C}}$ have the same matrix with respect to this basis. Thus the formula for $\det S$, which by definition equals $\det S_{\mathbb{C}}$, is given by Proposition 9.5. In particular, $\det S$ is real. The multiplicativity of \det on linear operators on a real vector space follows from the corresponding property on complex vector spaces and the multiplicativity of complexification: $(AB)_{\mathbb{C}} = A_{\mathbb{C}}B_{\mathbb{C}}$ whenever A and B are linear operators on a real vector space.

The tools we've developed provide a natural connection between determinants and volumes in \mathbb{R}^n . To understand that connection, first we need to explain what is meant by the square root of an operator times its adjoint. Suppose S is a linear operator on a real vector space U . If λ is an eigenvalue of S^*S and $u \in U$ is a corresponding non-zero eigenvector, then

$$\lambda \langle u, u \rangle = \langle \lambda u, u \rangle = \langle S^*Su, u \rangle = \langle Su, Su \rangle,$$

and thus λ must be a non-negative number. Clearly, S^*S is self-adjoint, and so there is a basis of U consisting of eigenvectors of S^*S (by the real Spectral Theorem, which is Theorem 8.3). We can think of S^*S as a diagonal matrix with respect to this basis. The entries on the diagonal, namely the eigenvalues of S^*S , are all non-negative, as we have just seen. The *square root* of S^*S , denoted $\sqrt{S^*S}$, is the linear operator on U corresponding to the diagonal matrix obtained by taking the non-negative square root of each entry of the matrix of S^*S . Obviously, $\sqrt{S^*S}$ is self-adjoint, and its square equals S^*S . Also, the multiplicativity of \det shows that

$$(\det \sqrt{S^*S})^2 = \det(S^*S) = (\det S^*)(\det S) = (\det S)^2.$$

Thus $\det \sqrt{S^*S} = |\det S|$ (because $\det \sqrt{S^*S}$ must be non-negative).

The next lemma provides the tool we will use to reduce the question of volume change by a linear operator to the self-adjoint case. It is called the *polar decomposition* of an operator S , because it resembles the polar decomposition of a complex number $z = e^{i\theta}r$. Here r equals \sqrt{zz} (analogous to $\sqrt{S^*S}$ in the lemma), and multiplication by $e^{i\theta}$ is an isometry on \mathbb{C} (analogous to the isometric property of A in the lemma).

Lemma 9.6. *Let S be a linear operator on a real inner product space U . Then there exists a linear isometry A on U such that $S = A\sqrt{S^*S}$.*

Proof: For $u \in U$ we have

$$\|\sqrt{S^*S}u\|^2 = \langle \sqrt{S^*S}u, \sqrt{S^*S}u \rangle = \langle S^*Su, u \rangle = \langle Su, Su \rangle = \|Su\|^2.$$

In other words, $\|\sqrt{S^*S}u\| = \|Su\|$. Thus the function A defined on $\text{ran } \sqrt{S^*S}$ by $A(\sqrt{S^*S}u) = Su$ is well defined and is a linear isometry from $\text{ran } \sqrt{S^*S}$ onto $\text{ran } S$. Extend A to a linear isometry of U onto U by first extending A to be any isometry of $(\text{ran } \sqrt{S^*S})^{\perp}$ onto $(\text{ran } S)^{\perp}$ (these two spaces have the same dimension, because we have just seen that there is a linear isometry of $\text{ran } \sqrt{S^*S}$ onto $\text{ran } S$), and then extend A to all of U by linearity (with the Pythagorean Theorem showing that A is an isometry on all of U). The construction of A shows that $S = A\sqrt{S^*S}$, as desired. \square

Now we are ready to give a clean, illuminating proof that a linear operator changes volumes by a factor of the absolute value of the determinant. We will not formally define volume, but only use the obvious properties that volume should satisfy. In particular, the subsets E of \mathbf{R}^n considered in the theorem below should be restricted to whatever class the reader uses most comfortably (polyhedrons, open sets, or measurable sets).

Theorem 9.7. *Let S be a linear operator on \mathbf{R}^n . Then*

$$\text{vol } S(E) = |\det S| \text{vol } E$$

for $E \subset \mathbf{R}^n$.

Proof: Let $S = A\sqrt{S^*S}$ be the polar decomposition of S as given by Lemma 9.6. Let $E \subset \mathbf{R}^n$. Because A is an isometry, it does not change volumes. Thus

$$\text{vol } S(E) = \text{vol } A(\sqrt{S^*S}(E)) = \text{vol } \sqrt{S^*S}(E).$$

But $\sqrt{S^*S}$ is self-adjoint, and we already noted at the beginning of this section that each self-adjoint operator changes volume by a factor equal to the absolute value of the determinant. Thus we have

$$\text{vol } S(E) = \text{vol } \sqrt{S^*S}(E) = |\det \sqrt{S^*S}| \text{vol } E = |\det S| \text{vol } E,$$

as desired. □

10. CONCLUSION. As mathematicians, we often read a nice new proof of a known theorem, enjoy the different approach, but continue to derive our internal understanding from the method we originally learned. This paper aims to change drastically the way mathematicians think about and teach crucial aspects of linear algebra. The simple proof of the existence of eigenvalues given in Theorem 2.1 should be the one imprinted in our minds, written on our blackboards, and published in our textbooks. Generalized eigenvectors should become a central tool for the understanding of linear operators. As we have seen, their use leads to natural definitions of multiplicity and the characteristic polynomial. Every mathematician and every linear algebra student should at least remember that the generalized eigenvectors of an operator always span the domain (Proposition 3.4)—this crucial result leads to easy proofs of upper-triangular form (Theorem 6.2) and the Spectral Theorem (Theorems 7.5 and 8.3).

Determinants appear in many proofs not discussed here. If you scrutinize such proofs, you'll often discover better alternatives without determinants. Down with Determinants!

*Department of Mathematics
Michigan State University
East Lansing, MI 48824
axler@math.msu.edu*

Answer to Picture Puzzle
(p. 138)
Mary Ellen Rudin, in 1970.

NOTES

Edited by: John Duncan

Where Not to Find the Critical Points of a Polynomial—Variation on a Putnam Theme

Peter Andrews

If a real polynomial of degree n has distinct roots $r_1 < r_2 < \cdots < r_n$ then Rolle's Theorem tells us that the critical points x_i , $i = 1, 2, \dots, n - 1$ lie in the intervals (r_i, r_{i+1}) respectively. This note addresses the question of what restrictions there might be on where in these intervals the critical points can be located? Two recent notes in this Monthly, for instance, show that the critical points cannot be any closer together than the roots themselves [4], [1].

Along these same lines, Problem A-3 on the 1991 Putnam competition reads:

Find all real polynomials $p(x)$ of degree $n \geq 2$ for which there exist real numbers $r_1 < r_2 < \cdots < r_n$ such that

1. $p(r_i) = 0$, $i = 1, 2, \dots, n$ and
2. $p'(\frac{r_i + r_{i+1}}{2}) = 0$, $i = 1, 2, \dots, n - 1$,

where $p'(x)$ is the derivative of $p(x)$.

This essentially asks, "Can the roots r_i be chosen so that *each* critical point x_i is at the midpoint of the interval (r_i, r_{i+1}) ?" When $n > 2$ the answer is no!

Since $p(x)$ is a degree- n polynomial with n distinct real roots, it can be written (at least up to a constant multiple) as $p(x) = \prod_{i=1}^n (x - r_i)$. The product rule and a straightforward calculation show that $p'((r_1 + r_2)/2) = 0$ (and likewise $p'((r_{n-1} + r_n)/2) = 0$) if and only if $n = 2$ [3, p. 719–720].

This tells us that, unless $n = 2$, neither the leftmost nor the rightmost critical point can be exactly halfway between the surrounding roots. A closer study reveals that all the critical points are bounded away from the roots in a rather intriguing manner.

Theorem. If $p(x)$ is a degree- n , real polynomial with distinct roots $r_1 < r_2 < \cdots < r_n$ and critical points $x_1 < x_2 < \cdots < x_{n-1}$ then

$$\frac{1}{n - i + 1} < \frac{x_i - r_i}{r_{i+1} - r_i} < \frac{i}{i + 1}.$$

The main observation we need is the "Root Dragging Theorem" of [1]. It states that if we move *any* of the roots to the right (i.e. increase them) then *all* of the critical points move to the right. We will sketch an alternative proof to that of [1].

If $p(x) = \prod_{i=1}^n (x - r_i)$, then since $p'(x_i) = 0$, x_i must satisfy the equation

$$x_i - r_i = - \frac{\prod_{j \neq i} (x_i - r_j)}{\sum_{k \neq i} \prod_{j \neq k, i} (x_i - r_j)}. \quad (1)$$

If we define $u_j = x_i - r_j$ and $F_i(u_1, \dots, u_n) = -\prod_{j \neq i} u_j / \sum_{k \neq i} \prod_{j \neq k, i} u_j$, then (1) becomes

$$u_i = F_i(u_1, \dots, u_n). \quad (2)$$

Moreover, $\partial F_i / \partial u_j < 0$ for $i \neq j$. This means that if we increase r_j ($j \neq i$) then u_j decreases and F_i increases. This makes $u_i < F_i(u_1, \dots, u_n)$. If x_i is now decreased then *each* u_j will decrease and this inequality will get even worse. Thus, the only way we can get back to equality in (2) and hence to the critical point is by moving x_i to the right as well.

Now, to return to the proof of our theorem, suppose we fix r_i and r_{i+1} . We can move x_i as far to the right as possible by letting $r_1, \dots, r_{i-1} \rightarrow r_i$ and $r_{i+2}, \dots, r_n \rightarrow \infty$. This suggests that we look at

$$q_b(x) = (x - r_i)^i (x - r_{i+1})(x - b)^{n-i-1}, \quad (3)$$

and let $b \rightarrow \infty$.

Differentiating $q_b(x)$ shows us that the critical point x_i must be the leftmost root of the quadratic equation

$$\begin{aligned} nx^2 - ((n-1)r_{i+1} + (n-i)r_i + (i+1)b)x \\ + (ibr_{i+1} + br_i + (n-i-1)r_{i+1}r_i) = 0. \end{aligned} \quad (4)$$

As $b \rightarrow \infty$, x_i then approaches the leftmost (and only) root of the linear equation

$$-(i+1)x + ir_{i+1} + r_i = 0. \quad (5)$$

Thus, $x_i \nearrow (ir_{i+1} + r_i)/(i+1)$ and $(x_i - r_i)/(r_{i+1} - r_i) \nearrow i/(i+1)$.

Similarly, to see how far to the left in (r_i, r_{i+1}) and x_i could be, we look at

$$s_b(x) = (x + b)^{i-1} (x - r_i)(x - r_{i+1})^{n-i} \quad (6)$$

and let $b \rightarrow \infty$. This time, x_i decreases toward the root of the equation

$$(n-i+1)x - r_{i+1} - (n-i)r_i = 0. \quad (7)$$

This means that $x_i \searrow (r_{i+1} + (n-i)r_i)/(n-i+1)$ and $(x_i - r_i)/(r_{i+1} - r_i) \searrow 1/(n-i+1)$.

Notice that we are measuring the position of x_i in the interval (r_i, r_{i+1}) by looking at $\sigma_i = (x_i - r_i)/(r_{i+1} - r_i)$, the ratio in which x_i divides (r_i, r_{i+1}) . If we define, for each polynomial $p(x) = \prod_{i=1}^n (x - r_i)$ with $r_1 < r_2 < \dots < r_n$, the $(n-1)$ -tuple of these ratios, $\sigma_p = (\sigma_1, \sigma_2, \dots, \sigma_{n-1})$, and for each $n > 2$, the set

$$X_n = \prod_{i=1}^{n-1} (1/(n-i+1), i/(i+1)), \quad (8)$$

then our theorem can be re-stated as saying that

$$\sigma_p \in X_n \quad (9)$$

where $n > 2$ is the degree of p .

It is not clear, however, that every $(\sigma_1, \sigma_2, \dots, \sigma_{n-1}) \in X_n$ can be realized as the ratio vector of a degree- n polynomial with distinct roots. Suppose we define $Y_n \subseteq X_n$ to be the set of those $(n-1)$ -tuples that are the ratio vectors of

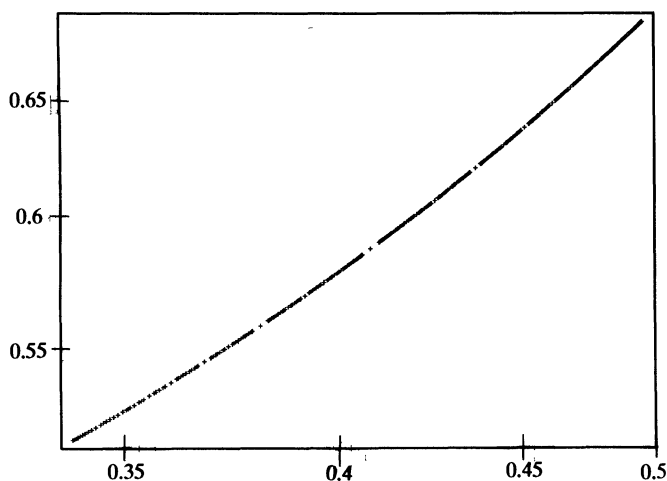


Figure 1

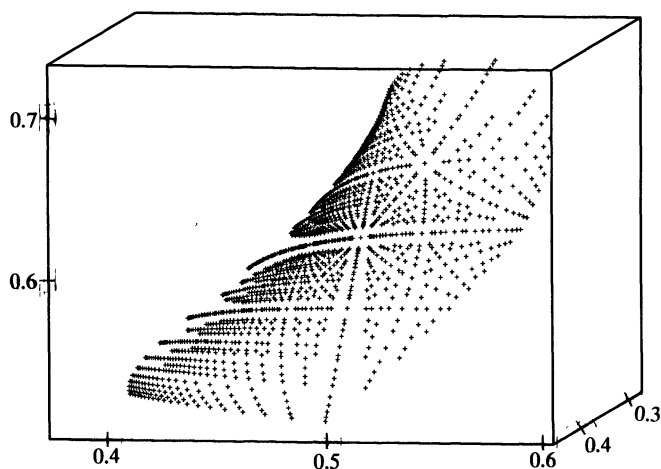


Figure 2

polynomials. Now our theorem appears as

$$Y_n \subseteq X_n \quad (10)$$

and the essence of the 1991 Putnam question could be re-phrased as:

For which values of $n > 2$ is the $(n - 1)$ -tuple $(1/2, 1/2, \dots, 1/2)$ in Y_n ?

Since $(1/2, 1/2, \dots, 1/2) \notin X_n$, (10) shows that it can't be in Y_n .

This still leaves the question of whether or not $Y_n = X_n$ and, if not, exactly what does Y_n look like? While we cannot answer these questions yet, we can leave you with some pictorial evidence that Y_n is a rather interesting hypersurface in X_n .

Since the location of the critical points relative to the roots is clearly invariant under translation, we can assume that the smallest root, r_1 , is at the origin. Figure 1 was produced by plotting σ_p for the cubic polynomials $p(x) = x(x - r_2)(x - r_3)$, as r_2 ranges from 0.1 to 2.0 in steps of 0.1 and r_3 ranges from $r_2 + 0.1$ to 4.0 in

steps of 0.1. This picture inspires one to look for a single equation satisfied by σ_1 and σ_2 . In fact, it is not too hard to show that

$$(1 - \sigma_1)\sigma_2 = \frac{1}{3}. \quad (11)$$

To see this, translate p so that $r_2 = 0$. Then $p'(x) = 3x^2 - 2(r_1 + r_3)x + r_1r_3$. From this, we see that the product of the two roots of p' is $r_1r_3/3$. However, the roots of p' are $(1 - \sigma_1)r_1$ and σ_2r_3 .¹

Figure 2 was produced in a similar manner but using quartic polynomials of the form $p(x) = x(x - r_2)(x - r_3)(x - r_4)$. The critical points were approximated by numerically solving the cubic equation $p'(x) = 0$ using Maple's **fsolve** procedure. This time Y_4 clearly appears to be a smooth surface in X_4 .

REFERENCES

1. B. Anderson, Polynomial root dragging, *Amer. Math. Monthly* 100 (1993) 864–866.
2. R. Gelca, A short proof of a result on polynomials, *Amer. Math. Monthly* 100 (1993) 936–937.
3. L. F. Klosinski, G. L. Alexanderson and L. C. Larson, The fifty-second William Lowell Putnam mathematical competition, *Amer. Math. Monthly* 99 (1992), 715–724.
4. P. Walker, Separation of the zeros of polynomials, *Amer. Math. Monthly* 100 (1993) 272–273.

Department of Mathematics
Eastern Illinois University
Charleston, IL 61920
cfpga@eiu.edu

A Short Path to the Shortest Path

Peter D. Lax

This note contains a demonstration of the isoperimetric inequality. Our proof is somewhat simpler and more straightforward than the usual ones; it is eminently suitable for presentation in an honors calculus course.

1. *The Isoperimetric Inequality* says that a closed plane curve of length 2π encloses an area $\leq \pi$. Equality holds only for a circle.

Let $x(s)$, $y(s)$ be the parametric presentation of the curve, s arclength, $0 \leq s \leq 2\pi$. Suppose that we have so positioned the curve that the points $x(0)$, $y(0)$ and $x(\pi)$, $y(\pi)$ lie on the x -axis, i.e.

$$y(0) = 0 = y(\pi). \quad (1)$$

The area enclosed by the curve is given by the formula

$$A = \int_0^{2\pi} y\dot{x} ds, \quad (2)$$

where the dot $\dot{}$ denotes differentiation with respect to s . We write this integral as the sum $A_1 + A_2$ of an integral from 0 to π and from π to 2π , and show that each is $\leq \frac{\pi}{2}$.

¹The author thanks the referee for this particularly nice derivation of (11).

According to a basic inequality,

$$ab \leq \frac{a^2 + b^2}{2};$$

equality holds only when $a = b$. Applying this to $y = a$, $\dot{x} = b$, we get

$$A_1 = \int_0^\pi y \dot{x} ds \leq \frac{1}{2} \int_0^\pi (y^2 + \dot{x}^2) ds. \quad (3)$$

Since s is arclength, $\dot{x}^2 + \dot{y}^2 = 1$; so we can rewrite (3) as

$$A_1 \leq \frac{1}{2} \int_0^\pi (y^2 + 1 - \dot{y}^2) ds. \quad (3')$$

Since $y = 0$ at $s = 0$ and π , we can factor y as

$$y(s) = u(s) \sin s, \quad (4)$$

u bounded and differentiable. Differentiate (4):

$$\dot{y} = \dot{u} \sin s + u \cos s.$$

Setting this into (3') gives

$$A_1 \leq \frac{1}{2} \int_0^\pi [u^2(\sin^2 s - \cos^2 s) - 2u\dot{u} \sin s \cos s - \dot{u}^2 \sin^2 s + 1] ds. \quad (5)$$

The product $2u\dot{u}$ is the derivative of u^2 ; integrating by parts changes (5) into

$$A_1 \leq \frac{1}{2} \int_0^\pi (1 - \dot{u}^2 \sin^2 s) ds,$$

clearly $\leq \pi/2$. Equality holds only if $\dot{u} \equiv 0$, which makes $y(s) \equiv \text{constant} \sin s$. Since equality in (3) holds only if $y = \dot{x} = \sqrt{1 - \dot{y}^2}$, $y(s) \equiv \pm \sin s$, $x(s) \equiv \mp \cos s + \text{constant}$. This is a semicircle. Q.e.d.

*Courant Institute of Mathematical Sciences
New York University
251 Mercer Street
New York, NY 10012*

A Note on Entire Solutions of the Eiconal Equation

Dmitry Khavinson

The eiconal equation $\sum_{i=1}^n (\partial u / \partial x_i)^2 = 1$, $u: \mathbf{R}^n \rightarrow \mathbf{R}$ is one of the main equations of geometrical optics. Its characteristics represent the light rays, while the level surfaces of solution u can be thought of as wave fronts (cf., e.g., [3]). Here,

however, we are interested in the complex-analytic solutions of that equation. Consider a holomorphic solution u of the eiconal equation in two variables

$$u_x^2 + u_y^2 = 1, \quad ((x, y) \in \mathbb{C}^2). \quad (1)$$

The purpose of this note is to prove the following:

Theorem 1. *Let u be an entire solution of equation (1) in \mathbb{C}^2 . Then u is a linear function, i.e., $u = ax + by + c$, where $a, b, c \in \mathbb{C}$, and $a^2 + b^2 = 1$.*

Note. The theorem fails in all dimensions ≥ 3 . For example, consider in \mathbb{C}^3 the function $u(x, y, z) := x + f(y + iz)$, where f is an arbitrary entire function of one variable. Then, u is obviously nonlinear and satisfies the eiconal equation

$$u_x^2 + u_y^2 + u_z^2 = 1$$

everywhere in \mathbb{C}^3 .

Proof of the theorem. Changing variables to $z = x + iy$, $w = x - iy$, we reduce the problem to showing that an entire solution $u(z, w)$ satisfying

$$u_z \cdot u_w = 1 \quad (2)$$

is linear. Since $u_z \neq 0$ in \mathbb{C}^2 , $u_z = e^\gamma$, while $u_w = e^{-\gamma}$, where $\gamma(z, w)$ is entire. We have

$$u_{zw} = \gamma_w e^\gamma = -\gamma_z e^{-\gamma},$$

or

$$\gamma_w u_z + \gamma_z u_w \equiv 0. \quad (3)$$

Consider a level surface $\gamma_c := \{(z, w): \gamma(z, w) = c\}$, $c \in \mathbb{C}$. By (3), on γ_c

$$-\frac{\gamma_z}{\gamma_w} = \frac{u_z}{u_w} = e^{2c} = \text{const}. \quad (4)$$

But the left-hand side in (4) is the derivative of the implicit function $f(z)$ defined by

$$\gamma(z, f(z)) = c.$$

Thus, $f'(z) = e^{2c} = \text{const}$, and so $\gamma_c = \{w = f(z)\}$ is a complex line with the “slope” e^{2c} . Moreover, $\gamma_{c_1} \cap \gamma_{c_2} = \emptyset$ whenever $c_1 \neq c_2$. Otherwise, γ , which is an entire function, would be taking two different values at the point of intersection of γ_{c_1} and γ_{c_2} . Hence, all “slopes” of the γ_c ’s must be the same, i.e., $e^{2\gamma} \equiv \text{const}$, and so, $\gamma \equiv \text{const}$, and

$$u_z = e^\gamma = \text{const}, \quad u_w = e^{-\gamma} = \text{const}.$$

Thus, $u = Az + Bw + C$, with $AB = 1$ as claimed.

Remarks

(i) Peculiar cylinders $\Gamma := \{(x, y, z): x + f(y + iz) = \text{const}\}$, in \mathbb{C}^3 that prevent extending the theorem to higher dimensions were first noted by G. Johnsson [4] in connection with the Cauchy problem for the Laplace equation (also, cf. [6]).

(ii) Professor Tabachnikov has kindly pointed out to the author, that Theorem 1 does extend to higher dimensions under an additional assumption that the restriction of the function u to \mathbf{R}^n is real-valued. In that case, the level surfaces of function u represent the fronts of light rays moving along normals to a given surface, say $\Gamma_0 := \{x \in \mathbf{R}^n: u(x) = 0\}$, with a constant speed. The eiconal equation then implies that neither of those level surfaces have singular points, i.e., all normals to Γ_0 do not contain any finite focal points. Hence, Γ_0 has zero curvature (cf., e.g., [9, Section 6]). So, all level surfaces of u are parallel planes, and then, e.g., by solving explicitly the initial value problem for the eiconal equation, one readily verifies that u must be a linear function.

(iii) Differentiating (2) with respect to z and w once more, we obtain, after some straightforward algebraic manipulations, that u satisfies

$$u_{zz}u_{ww} - u_{zw}^2 = 0. \quad (5)$$

(5) is a degenerate Monge-Ampère equation. There is an enormous literature dedicated to the study of such non-degenerate equations, i.e., with a non-vanishing right-hand side, e.g., cf., [3], [7], and the references cited there. Some of the results stemming from a celebrated theorem of S. Bernstein [1, 2, 5, 8, 10, 11, 12], that a minimal surface over a whole plane must be a plane, seem to be very close in flavor to Thm. 1. For example, a theorem of Jörgens [5] states that a C^2 -solution in \mathbf{R}^2 of (5) with the right-hand side equal to 1 must be a quadratic polynomial. (For another proof of that based on ideas from [7], see [10].) Perhaps, one could revise some of those arguments to include the degenerate case (5), and then in view of (2), obtain another high-ground proof of Thm. 1.

For this last remark, and the references, I am indebted to Professor H. S. Shapiro.

REFERENCES

1. S. Bernstein, Über ein geometrisches Theorem und seine Anwendung auf die partiellen Differentialgleichungen vom elliptischen Typus, *Math. Zeit.* 26 (1927), 551–558.
2. L. Bers, Isolated singularities of minimal surfaces, *Ann. Math.* 53 (1951), 364–380.
3. R. Courant, and D. Hilbert, *Methods of Mathematical Physics, II*, Wiley, 1989.
4. G. Johnsson, The Cauchy problem in \mathbf{C}^n for linear second-order partial differential equations with data on a quadratic surface, *Mem. Amer. Math. Soc.*, to appear.
5. K. Jörgens, Über die Lösungen der Differentialgleichung $rt - s^2 = 1$, *Math. Ann.* 127 (1954), 130–134.
6. D. Khavinson, Singularities of harmonic functions in \mathbf{C}^n , *Proc. Symp. Pure and Appl. Math., A.M.S.* 52 (1991), Part 3, 207–217.
7. H. Levy, A priori limitations for solutions of Monge-Ampère equations, I and II, *Trans. Amer. Math. Soc.* 37 (1935), 417–434, and 41 (1937), 365–374.
8. E. J. Mickle, A remark on a theorem of Serge Bernstein, *Proc. Amer. Math. Soc.* 1 (1950), 86–89.
9. J. Milnor, *Morse Theory*, Princeton University Press, Princeton, New Jersey, 1963.
10. J. C. C. Nitsche, Elementary proof of Bernstein's theorem on minimal surfaces, *Ann. Math.* 66 (1957), 593–594.
11. J. C. C. Nitsche, *Lectures on Minimal Surfaces*, vol. 1, Cambridge University Press, 1989.
12. T. Rado, Zu einem Satze von S. Bernstein über Minimalflächen im Gromen, *Math. Zeit.* 26 (1927), 559–565.

Department of Mathematics
University of Arkansas
Fayetteville, Arkansas 72701
dk24653@uafsysb.uark.edu

The Uniqueness Aspect of the Fundamental Theorem of Finite Abelian Groups

David B. Surowski

We shall use additive notation for abelian groups. The following is well-known to every graduate student of mathematics:

Fundamental theorem of finite Abelian groups. *Let A be a finite abelian group. Then there exist cyclic subgroups Z_1, Z_2, \dots, Z_r of orders $m_1, m_2, \dots, m_r > 1$, respectively, satisfying $m_2 | m_1, m_3 | m_2, \dots, m_r | m_{r-1}$ such that*

$$A = Z_1 \oplus Z_2 \oplus \cdots \oplus Z_r.$$

Furthermore, the integers r and m_1, \dots, m_r are uniquely determined.

Note that the above theorem involves two parts: an existence part and a uniqueness part. While there are many short papers that provide novel proofs of the existence aspect, the uniqueness aspect has been largely neglected. Indeed, an alarming number of textbook treatments of the “Fundamental Theorem” do not even mention uniqueness as part of the theorem. Those treatments that do address uniqueness all, in varying degrees, obtain the uniqueness along the lines of the argument as given in Mac Lane and Birkoff’s standard text [2], or by an analysis of the i -rowed minors of the “relations” matrix defining A . (Compare [1; Theorem 3.9]; this amounts to a proof of the uniqueness, up to associates of the “Smith Canonical Form” of the relations matrix defining A .) There may be some who, on the grounds of “purity,” might object to arguments akin to those in [2], as they invoke the uniqueness of dimension of a vector space. While this is hardly a serious objection, our argument below is quite independent of even this simple result.

Thus, assume that we have decompositions of the finite abelian group A into direct sums of cyclic groups:

$$Z_1 \oplus Z_2 \oplus \cdots \oplus Z_r = A = Z'_1 \oplus Z'_2 \oplus \cdots \oplus Z'_s,$$

where $|Z_i| = m_i$, $|Z'_j| = m'_j$, $i = 1, \dots, r$, $j = 1, \dots, s$, and that the above divisibility conditions on the orders m_j, m'_j hold. Note first of all, that if we set $m = \exp(A)$ = least positive integer m such that $ma = 0$ for all $a \in A$, then $m_1 = m'_1 = m$. We call m the *exponent* of A . Thus, we have decompositions of the form

$$Z \oplus B = A = Z' \oplus B',$$

where Z, Z' are cyclic of order m . The idea is to prove that there exists an automorphism $\phi: A \rightarrow A$ such that $\phi(Z) = Z'$, for then it would follow that $B \cong B'$, and the desired uniqueness would follow by induction.

We hasten to concede a small logical glitch in the above induction. Indeed, we have started with a fixed abelian group A with two direct sum decompositions; after application of the above isomorphism we obtain two isomorphic, but not identical groups, viz., B and B' . However, this is not a serious problem and the student should have no difficulty in enunciating an induction hypothesis sufficiently general to be applicable here.

What is really going on is summarized in the following:

Theorem. *The group $\text{Aut}(A)$ of automorphisms of A acts transitively on the cyclic subgroups of order $m = \exp(A)$ of A .*

In other words, given any two cyclic subgroups $C, C' \leq A$, both of order m , then there exists an automorphism $\psi: A \rightarrow A$ with $\psi(C) = C'$. In fact, we'll prove a slightly stronger result, namely that $\text{Aut}(A)$ actually acts transitively on the elements of order m in A .

To prove this result, it suffices to prove that if p is a prime and if p^k is the highest power of p that divides m , then $\text{Aut}(A)$ acts transitively on the elements of order p^k in A . It therefore suffices to consider the case in which the exponent m is itself a prime power: $m = p^k$.

Thus we have a decomposition $A = Z_1 \oplus Z_2 \oplus \cdots \oplus Z_r$, where Z_i is cyclic of order p^{k_i} , and where $k_1 = k \geq k_2 \geq \cdots \geq k_r$.

The following two lemmas are very easy, but fundamental.

Lemma 1. *Let $B = B_1 \oplus B_2$ be an abelian group and let $\mu_1: B_1 \rightarrow B$, $\mu_2: B_2 \rightarrow B$ be injective homomorphisms. If $\mu_1(B_1) \cap \mu_2(B_2) = 0$, then the mapping $\mu: B \rightarrow B$ defined by $\mu(b_1 + b_2) = \mu_1(b_1) + \mu_2(b_2)$, $b_1 \in B_1$, $b_2 \in B_2$ is an automorphism of B .*

Lemma 2. *Let $A = Z_1 \oplus Z_2 \oplus \cdots \oplus Z_r$ be as above, and assume that $Z_i = \langle z_i \rangle$ is cyclic of order p^{k_i} , $i = 1, \dots, r$. Assume that $k = k_1 = k_2 = \cdots = k_l$, $k_{l+1} < k$. If*

$$a = \sum_{i=1}^r \alpha_i z_i \in A,$$

then a has order p^k if and only if $p \nmid \alpha_j$, for some j , $1 \leq j \leq l$.

Proof: Simply note that because $Z_1 \oplus Z_2 \oplus \cdots \oplus Z_r$ is direct sum, we conclude that the order of $\sum_{i=1}^r \alpha_i z_i$ is equal to the least common multiple of the individual orders $o(\alpha_i z_i)$, $i = 1, \dots, r$. Since $p \nmid \alpha_j$, we see that $o(\alpha_j z_j) = o(z_j) = p^k$.

Thus, let $a \in A$ be an arbitrary element of order p^k . We shall show that there exists an automorphism $\phi: A \rightarrow A$ such that $\phi(z_1) = a$. If we write

$$a = \sum_{i=1}^r \alpha_i z_i,$$

we may assume that $p \nmid \alpha_1$. Indeed, if $p \mid \alpha_j$, then an automorphism of A that interchanges Z_1 and Z_j will reduce us to this situation. Next, according to *Lemma 1*, write $A = Z_1 \oplus B_2$, where $B_2 = Z_2 \oplus \cdots \oplus Z_r$, and define injections $\mu_1: Z_1 \rightarrow A$, $\mu_2: B_2 \rightarrow A$ by setting $\mu_1(z_1) = a$, $\mu_2 = 1_{B_2}$. Therefore the map $\mu: A \rightarrow A$, given by $\mu(w_1 + b_2) = \mu_1(w_1) + \mu_2(b_2)$ defines an automorphism of A which carries z_1 to a .

REFERENCES

1. N. Jacobson, *Basic Algebra I*, Second edition, W. H. Freeman, New York, 1985.
2. S. MacLane and G. Birkhoff, *Algebra*, Macmillan, New York, 1968.

*Department of Mathematics
Kansas State University
Manhattan, KS 66506-2602
dbski@math.ksu.edu*

UNSOLVED PROBLEMS

Edited by: Richard Guy & Richard Nowakowski

In this department the MONTHLY presents easily stated unsolved problems dealing with notions ordinarily encountered in undergraduate mathematics. Each problem should be accompanied by relevant references (if any are known to the author) and by a brief description of known partial or related results. Typescripts should be sent to Richard Guy, Department of Mathematics & Statistics, The University of Calgary, Alberta, Canada T2N 1N4.

Coin-Weighing Problems

Richard K. Guy and Richard J. Nowakowski

The question of finding a single counterfeit coin from a set of regular coins in the fewest number of weighings using just a balance beam has been a notorious problem. The regular coins are all the same weight while the counterfeit coin is a different weight.

The problem was popular on both sides of the Atlantic during World War II ([14, 15, 20, 21, 27, 34, 39, 46]; it was even suggested that it should be dropped over Germany in an attempt to sabotage their war effort; see [35, 40, 43] for some history. In solving [39] Kaplansky, Neugebauer and Pennell gave the following general solution to the problem of underweight counterfeit coins: *If $3^{n-1} \leq N < 3^n$ then n weighings suffice to show if there is (and to identify) a counterfeit coin among N coins.* If it is known that a counterfeit coin exists then n weighings will identify the coin from among N coins if $3^{n-1} < N \leq 3^n$. Dyson [12] gave an elegant solution using ternary labels when it is not known if the counterfeit coin is heavy or light; see [11] for a solution in verse. In this case, n weighings suffice

- (i) if $N \leq (3^n - 3)/2$ and it is required to find if the dud is heavy or light;
- (ii) if $N \leq (3^n - 1)/2$, given an extra coin known to be good, and it is required to find if the dud is heavy or light; and
- (iii) $N \leq (3^n + 1)/2$ if there is a good coin but the relative weight of the counterfeit coin is not required.

In the solution to the general problem posed in [15], the editors note that all the solutions so far consider the coins to be distinguishable when in the balance pan. They show that if the coins in a scale pan are to be considered as a single set, then n weighings will find a coin amongst $N \leq (7 \times 3^{n-2} - 1)/2$.

There are many other variants [1, 7, 9, 32]. Forysthe, a responder to [14], seems to be the first to ask the question using a spring balance i.e. a weighing device that will return the exact weight; see also [33, 41]. Christen [8] asks the question for two

counterfeit coins but of complementary weights. Hwang [25] proposes and analyses many weighing schemes.

Shapiro's problem [41] assumes N coins, $N - 1$ of weight a and one of weight b where a and b are known, and, as with Forsythe, an accurate scale. He asks for the least number of weighings to determine which coin has weight b , where the weighing scheme must be given in advance. Söderberg and Shapiro [44] ask the more general question of how many weighings are needed to determine which of N coins are of weight a and which of weight b if the numbers of each are not known. Denote this number of weighings by $f(N)$ then they show that (i) $f(N) \geq N/\log_2(N + 1)$; (ii) $f(3^{m-1}(3 + m)) \leq 3^m$; (iii) $f(5^{m-1}(2m + 5)) \leq 5^m$; and that (iv) $f(N) = O(N/\ln N)$. In addition Erdős and Rényi [13] (and several others independently) show that

$$f(N) = \frac{N}{\log_4 N} + O\left(\frac{N \ln \ln N}{(\ln N)^2}\right)$$

Cantor and Mills [6] and Lindström [28, 29, 30] give explicit weighing schemes for $N = 2^{k-1}k$ (also see [1]). The result of Liu [31] is not as good as this.

Another variant is that of deciding which coins are counterfeit out of N coins but the number of weighings is fixed. The "Lower Slobbovian Counterfeiters" [17, 4, 24] and ApSimon's Mints problem [2, 23] are examples.

Some years ago Sir Alexander Oppenheim reminded us of yet another variant. It was perhaps first stated by Bellman and Gluss [3]: use a beam balance to find k counterfeit coins among N coins where all the counterfeit coins are of the same weight. Let $w(N, k)$ be the least number of weighings required to find the k lighter coins. It is easy to see that $w(N, k)$ is at least $\log_3 \binom{N}{k}$. Pyber [36] showed that if there were no more than m light coins then they could be identified in $\left\lceil \log_3 \binom{n}{m} \right\rceil + 15m$ weighings.

The case $k = 2$ has recently attracted attention [5, 25, 45]. Tošić gave weighing procedures which improved on those of Cairns and showed that the lower bound could be attained apart from one possible extra weighing. For example, with seven coins weigh 123 against 456. If they balance, weigh 1 against 2 and 4 against 5. If 123 are heavier, then 4567 contain two light coins which can be determined in two more weighings. In the special case $w(N = 3^m, 2) = 2m$, the extra weighing is never needed.

In which cases is $w(N, 2) = \left\lceil \log_3 \binom{N}{2} \right\rceil + 1$? Is $N = 13$ the first?

If there are three lighter coins, then there is no new problem until we get to $w(6, 3) = 3$ which was the subject of a problem in [10]. Oppenheim showed that $w(7, 3) = w(8, 3) = 4$: first weigh three coins against three. Nine coins require 5 weighings. Can the lighter coins be identified in five weighings if $N = 12$? 3^m coins can be sorted in $3m$ weighings; this can be improved by at most one weighing; when?

If $k = 4$ we know that $w(8, 4) = w(9, 4) = 5$; although $\binom{8}{4} = 70 < 3^4$ it is not possible to make a weighing among 8 coins with 4 light each of whose outcomes leave less than 26 possibilities and while $26 < 3^3$ they cannot be separated by three weighings. We can show that $w(3^m, 4) \leq 4m - 1$.

If $k = 5$, then $w(10, 5) = 6$ and we can show that $w(3^m, 5) \leq 5m$, but this can almost certainly be improved.

The problem of Söderberg and Shapiro, but using a beam balance in place of a spring balance is: given N coins which each weigh one of two weights, determine

the least number, $W(N)$, of weighings required to find which coins are of each weight. Of course, if all the coins are of the same weight, we won't be able to say which of the two possible weights they are. We see that $W(1) = 0$, $W(2) = 1$, $W(3) = 2$, $W(4) = 3$, $W(5) = 4$, $W(6) = 4$ and generally $W(N) \geq \lceil \log_3 2^N \rceil$. For which N is there equality?

Notice that we don't require that the whole weighing scheme be given in advance, as has been done in the more elegant solutions of the famous 12 coin problem. The subsequent weighings depend on the results of the previous ones. We could also ask for the minimum number of weighings, if these are all to be prescribed before the first weighing is made.

REFERENCES

1. Martin Aigner, *Combinatorial Search*, Wiley 1988; MR 89k:68021. Chapter 2, Weighing Problems.
2. H. ApSimon, *Mathematical Byways in Ayling, Beeling, and Ceiling*, Oxford University Press, 1984.
3. R. Bellman and B. Gluss, On various versions of the defective coin problem, *Inform. and Control*, 4 (1961) 118–151.
4. J. Braun, The counterfeiters of Lower Slobbovia, this MONTHLY, 61 (1954) 472–473. [But see corrections by Carlitz and Selfridge, 62 (1955) 40–41.]
5. S. Cairns, Balance scale sorting, this MONTHLY, 70 (1963) 136–148; MR 26 #4929.
6. D. G. Cantor and W. H. Mills, Determining a subset from certain combinatorial properties, *Canad. J. Math*, 18 (1966), 42–48.
7. G. Chang, F. Hwang and S. Lin, Testing with two defectives, *Discrete Appl. Math.*, 4 (1982) 97–102.
8. C. Christen, Optimal detection of two complementary coins, *SIAM J. Alg. Discrete Math.*, 4 (1983) 101–110.
9. Claude Christen and Frank K. Hwang, Detection of a defective coin with partial weight information, this MONTHLY, 91 (1984) 173–179.
10. S. N. Collings (editor), Puzzle corner, *Bull. Inst. Math. Appl*, 20 (1984), p. 94, Puzzle number 79, Unbalanced coins II; p. 126, Solution; p. 153, Puzzle number 81, A colourless corollary; pp. 184–185, Solution.
11. Blanche Descartes, The twelve coin problem, *Eureka*, 13 (1950) 7, 20.
12. Freeman J. Dyson, Note 1931—The problem of the pennies, *Math. Gaz.*, 30 (1946) 231–234.
13. P. Erdős and A. Rényi, On two problems of information theory, *Publ. Hung. Acad. Sci.*, 8 (1963), 241–254.
14. Donald Eves, Problem E712 The extended coin problem, *Amer. Math. Monthly*, 53 (1946) 156. Solutions, E. D. Schell and Joseph Rosenbaum, this MONTHLY, 54 (1947) 46–48.
15. Nathan J. Fine, Problem 4203—The generalized coin problem, this MONTHLY, 53 (1946) 278, solution, 54 (1947) 489–491.
16. James Fixx, *More Games for the Superintelligent*, Warner Books, New York, 1972, p. 88.
17. L. R. Ford, Problem E1096, this MONTHLY, 61 (1954) 46.
18. Kobon Fujimura, Another balance scale problem, *Recreational Math. Mag.*, 10 (1962) 34 and 11 (1962) 42.
19. Kobon Fujimura and J. A. H. Hunter, There's always a way, *Recreational Math. Mag.*, 6 (1961) 67; editorial solution, 7 (1962) 53.
20. R. L. Goodstein, Note 1845—Find the penny, *Math. Gaz.*, 29 (1945) 227–229. [Erroneous solution, purporting to find dud among $(3^n - 2n + 3)/2$ coins.] Editorial note, Note 1930—Addendum 30 (1946) 231, gives correct solution.
21. Howard D. Grossman, The twelve-coin problem, *Scripta Math.*, 11 (1945) 360–361.
22. Howard D. Grossman, Ternary epitaph on coin problem, *Scripta Math.*, 14 (1948) 69–71.
23. R. K. Guy and R. J. Nowakowski, Mints, this MONTHLY, 101 (1994) 358–359.
24. M. Hendy, The retrieval of the Lower Slobbovian Counterfeiters, this MONTHLY, 87 (1980) 200–201. [But see 62 (1955) 40–41.]
25. F. K. Hwang, A tale of two coins, this MONTHLY, 94 (1987) 121–129.
26. K. Itkin, A generalization of the twelve-coin problem, *Scripta Math.*, 14 (1948) 67–68.
27. V. Karapetoff, The nine coin problem and the mathematics of sorting, *Scripta Math.*, 11 (1945) 186–187.

28. B. Lindström, On a combinatory detection problem I, *Magyar Tud. Akad. Mat. Kutató. Int. Közl.*, 9 (1964) 195–207; *MR* 29 #5750.
29. B. Lindström, On a combinatorial problem in number theory, *Canad. Math. Bull.*, 8 (1965), 477–490; *MR* 31 #5833.
30. B. Lindström, On a combinatory detection problem II, *Publ. Hung. Acad. Sci.*, 1 (1966) 353–361; *MR* 35 #4115.
31. Liu Teng-Sun, To weigh $5 + 2n$ coins of two different weights in $4 + n$ times, *J. Tianjin Univ.*, 1986 no. 4, 77–85; *MR* 88g:05006.
32. Bennet Manvel, Counterfeit coin problems, *Math. Mag.*, 50 (1977) 90–92; *MR* 55 #7792.
33. J. G. Mauldon, Problem E3023, this MONTHLY, 90 (1983) 645. Various solutions, 96 (1989) 254–258.
34. A. M. Mood, On Hotelling's weighing problem, *Ann. Math. Statist.*, 17 (1946) 432–446.
35. E. V. Newbery, The penny problem, Note 2342, *Math. Gaz.*, 37 (1953) 130.
36. L. Pyber, How to find many counterfeit coins, *Graphs Combin.*, 2 (1986), 173–177; *MR* 89c:05003.
37. C. W. Raine, Another approach to the twelve-coin problem, *Scripta Math.*, 14 (1948) 66–67.
38. J. A. Robertson, Those twelve coins again, *Scripta Math.*, 16 (1950) 111–115.
39. E. D. Schell, Problem E651—Weighed and found wanting, this MONTHLY, 52 (1945) 42. Solution, M. Dernham, 52 (1945) 397.
40. Benjamin L. Schwartz, Letter: Truth about false coins, *Math. Mag.*, 51 (1978) 254. [States that Schell told Michael Goldberg in 1945 that he had originated the problem.]
41. Harold S. Shapiro, Problem 1399—Counterfeit coins, this MONTHLY, 67 (1960) 82; Solution, Nathan J. Fine, 67 (1960) 697–698.
42. Cedric Austen Bardell Smith, The counterfeit coin problem, *Math. Gaz.*, 31 (1947) 31–39.
43. Dwight A. Stewart, The counterfeit coin, proposed in L. A. Graham, *Ingenious Mathematical Problems and Methods*, Dover, 1959, pp. 37–38. Solutions, D. B. Parkinson and Lester H. Green, pp. 196–198. [Problem appeared in the *Graham Dial*, October 1945.]
44. Staffan Söderberg and H. S. Shapiro, A combinatory detection problem, this MONTHLY, 70 (1963), 1066–1070.
45. R. Tošić, Two counterfeit coins, *Discrete Math.*, 46 (1983) 295–298.
46. Lothrop Withington, Another solution of the 12-coin problem, *Scripta Math.*, 11 (1945) 361–363.

Department of Mathematics
The University of Calgary
Calgary, Alberta
CANADA T2N 1N4

Department of Mathematics
Dalhousie University
Halifax, Nova Scotia
CANADA B3H 3J5

Bridges would not be safer if only people who knew
the proper definition of a real number were allowed to
design them.

—*Norman David Mermin (1935–)*
Topological Theory of Defects in Review of Modern Physics
July 1979 51, No. 3.

PROBLEMS AND SOLUTIONS

Edited by:

Richard T. Bumby, Fred Kochman and Douglas B. West

Proposed problems should be sent to the MONTHLY PROBLEMS address given on the inside front cover. Please include solutions, relevant references, etc. Three copies are requested.

Solutions of published problems should arrive before July 31, 1995 at the MONTHLY PROBLEMS address given on the inside front cover. Solutions should be typed with double spacing, including the problem number and the solver's name and mailing address. Two copies suffice. A self-addressed postcard or label should be included if an acknowledgement is desired.

An asterisk () after the number of a problem, or part of a problem, indicates that no solution is currently available. Partial solutions will be useful in such cases. Otherwise, the published solution is likely to be based on a solution which is complete and correct. Of course, an elegant partial solution or a method leading to a more general result is always useful and welcome. In addition, references to other appearances of MONTHLY problems or to solutions of these problems in the literature are also solicited.*

PROBLEMS

10431. *Proposed by Yury J. Ionin, Central Michigan University, Mt. Pleasant, MI.*

For positive integers n and s with $n \geq s$, the *falling factorial* $(n)_s$ is defined as $\frac{n!}{(n-s)!}$. Let $d(n, s)$ denote the greatest common divisor of the falling factorials $(n)_s$ and $(n+s)_s$.

Prove that $d(n, s) \mid (2s-1)_{\lfloor 4s/3 \rfloor}$.

10432. *Proposed by David M. Bloom, Brooklyn College, CUNY, Brooklyn, NY.*

Let

$$P = \{p \in \mathbb{Z}^+ : p \text{ is prime and } p \equiv 3 \pmod{4}\}.$$

For $p \in P$, let $S(p)$ denote the sum of all quadratic residues $(\text{mod } p)$ that lie in the interval $(0, p/2)$, and let $R(p)$ denote the least positive residue of $S(p) \pmod{p}$.

(a) Prove that R is one-to-one.

(b) Show that there are infinitely many positive integers that are not in the range of R .

10433. Proposed by Daniel R. L. Brown (student), Kenneth R. Davidson, and Jeffrey Shallit, University of Waterloo, Waterloo, Ontario, Canada.

Let x_1, x_2, x_3, \dots be any sequence of positive real numbers, and let k be any positive integer.

(a) Show that

$$\limsup_{n \rightarrow \infty} \frac{x_1 + x_2 + \dots + x_{n+1}}{x_n} \geq 4.$$

(b) More generally, show that

$$\limsup_{n \rightarrow \infty} \frac{x_1 + x_2 + \dots + x_{n+k}}{x_n} \geq \frac{(k+1)^{k+1}}{k^k}.$$

(c) Show that these bounds are best possible.

10434. Proposed by Daniel Goffinet, Saint Étienne, France.

Let P be the set of nonconstant periodic mappings from \mathbb{R} to \mathbb{R} , endowed with the topology derived from the supremum norm. Find the components of P .

10435. Proposed by Jonathan L. King, University of Florida, Gainesville, FL.

Let the function $K(x, y)$ be nonnegative and continuous on $0 \leq x \leq 1, 0 \leq y \leq 1$. Suppose that there are functions $f(x), g(x)$, each positive and continuous for $0 \leq x \leq 1$, such that

$$\int_0^1 f(y)K(x, y) dy = g(x) \text{ and } \int_0^1 g(y)K(x, y) dy = f(x).$$

Is it true that $f(x) = g(x)$ for $0 \leq x \leq 1$?

10436. Proposed by Donald A. Darling, Newport Beach, CA.

Let the unit interval $(0, 1)$ be divided at random into two subintervals. That is, form the intervals $L_1 = (0, X)$ and $R_1 = (X, 1)$ where X is a random variable uniformly distributed in $(0, 1)$. The interval L_1 is similarly divided, independently of the first division, into L_2 and R_2 , and the process of dividing the leftmost subinterval is continued indefinitely yielding two sequences $\{L_1, L_2, \dots\}$ and $\{R_1, R_2, \dots\}$ of intervals. Let the length of an interval I be denoted by $|I|$. Find the distribution of $S = \sum |L_n|$, i. e., the function $F(x) = \Pr \{S \leq x\}$.

10437. Proposed by J. Maurice Rojas (student), University of California, Berkeley, CA, and A T & T Bell Laboratories, Naperville, IL.

Let R be a ring (whose multiplication is not necessarily commutative or associative) without zero divisors. Let x_1, \dots, x_n be algebraically independent indeterminates over R which commute and associate amongst themselves and commute with the elements of R . Also assume the associative law for products of one element of R and two x_i . Prove the following.

(a) If $f \in R[x_1, \dots, x_n]$ is homogeneous, then any divisor of f is homogeneous.

(b) If $\alpha_1, \dots, \alpha_n$ are nonzero elements of R and d_1, \dots, d_n are nonnegative integers with $\gcd(d_1, \dots, d_n) = 1$, then the polynomial

$$\alpha_1 x_1^{d_1} + \dots + \alpha_n x_n^{d_n}$$

is irreducible in $R[x_1, \dots, x_n]$, i. e., every factorization has at most one nonconstant factor.

NOTES

(10432) A number a is a quadratic residue modulo p if $p \nmid a$ and a is congruent (mod p) to the square of an integer. Thus, for $p = 7$, the quadratic residues are 1, 2, -3 , so $R(7) = S(7) = 3$. For $p = 11$, the quadratic residues are 1, -2 , 3, 4, 5. Thus $S(11) = 13$ and $R(11) = 2$. (10435) If $K(x, y)$ is required to be positive, the conclusion is true. This positive result was problem B-4 of the 1993 Putnam Competition. (10436) It is easily shown that the related expression $\sum |R_n|$ is equal to 1 with probability one.

SOLUTIONS

A Trigonometric Matrix Norm

E 3473 [1991,956]. *Proposed by Lawrence J. Wallen, University of Hawaii at Manoa, Honolulu.*

Suppose $0 \leq \theta_1 < \theta_2 < \cdots < \theta_n < \pi$. Let A be the n by n matrix whose entry in the i th row and j th column is $\sin |\theta_i - \theta_j|$. Show that

$$\|A\| \leq \cot(\pi/2n)$$

and that the estimate is best possible. Here $\|A\|$ is defined as $\sup |AX|$, where the supremum is taken over all column vectors X in \mathbb{R}^n with Euclidean norm 1.

Solution I by the proposer. Since A has nonnegative entries with positive off-diagonal elements, A^2 has positive entries (except for $n = 2$). Perron-Frobenius Theory assures us that if ξ is the maximum modulus of the eigenvalues for A , then ξ is an eigenvalue of A with an eigenvector having strictly positive entries λ_i . Since A is symmetric, $\xi = \|A\|$.

Let P be a $2n$ -gon whose opposite sides σ_i and σ_{i+n} are parallel and of equal length λ_i ($i = 1, \dots, n$), and let edge σ_i , if extended, make angle θ_i with the x -axis.

Claim. Any such polygon P can be tessellated with $\binom{n}{2}$ parallelograms whose edges are translates of the pairs σ_i, σ_j for $1 \leq i < j \leq n$.

Proof. The result is clearly true for $n = 2$. This will be the basis of our induction. (The case $n = 1$ may also be considered to be vacuously true.)

Delete edges σ_n and σ_{2n} , and translate one of the two broken half-perimeters by σ_n to meet the other. This forms a $2(n-1)$ -gon P' that also has opposite sides parallel and of equal length. We may assume by induction that P' has the desired tessellation property. Also, the region swept out by the moving half-perimeter is tessellated by parallelograms whose edges are translates of the pairs (σ_i, σ_n) for $1 \leq i < n$. Putting these two observations together gives the required tessellation of P .

This dissection of P shows that

$$\text{Area}(P) = \sum_{i < j} \lambda_i \lambda_j \sin(\theta_j - \theta_i) = \frac{1}{2} \sum_{i, j} \lambda_i \lambda_j \sin |\theta_j - \theta_i| = \frac{\xi}{2} \left(\sum \lambda_i^2 \right).$$

Now, the isoperimetric quotient (area/perimeter²) of a convex m -gon is greatest for the regular m -gon, and this value is $(1/4m) \cot(\pi/m)$ (see I. M. Yaglom & V. G. Boltyansky, *Convex Figures*, Holt, Rinehart and Winston, 1961). Hence we have, by Cauchy's inequality

$$\frac{\xi}{2} \left(\sum \lambda_i^2 \right) \leq \left(2 \sum \lambda_i \right)^2 (8n)^{-1} \cot \left(\frac{\pi}{2n} \right) \leq (2n)^{-1} \left(n \sum \lambda_i^2 \right) \cot \left(\frac{\pi}{2n} \right),$$

giving the required bound.

If $\theta_k = k\pi/n$, $k = 0, \dots, n-1$, then it is easy to compute that all row sums of A are equal to

$$\sum_{k=1}^{n-1} \sin \left(\frac{k\pi}{n} \right) = \cot \left(\frac{\pi}{2n} \right).$$

Hence $\cot(\pi/2n)$ is an eigenvalue with $(1, 1, \dots, 1)^T$ as an eigenvector. Thus, the upper bound is attained.

Solution II by Anchorage Math Solutions Group, University of Alaska, Anchorage, AK. An example showing that the upper bound is attained was given in Solution I. It need not be repeated.

Write a_i for $|\theta_{i+1} - \theta_i|$. Then A is the real part of an Hermitian matrix $H = -iC$, where

$$C = \begin{bmatrix} 0 & e^{ia_1} & e^{i(a_1+a_2)} & \dots & e^{i(a_1+\dots+a_{n-1})} \\ -e^{-ia_1} & 0 & e^{ia_2} & \dots & e^{i(a_2+\dots+a_{n-1})} \\ -e^{-i(a_1+a_2)} & -e^{-ia_2} & 0 & \dots & e^{i(a_3+\dots+a_{n-1})} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -e^{-i(a_1+\dots+a_{n-1})} & -e^{-i(a_2+\dots+a_{n-1})} & -e^{-i(a_3+\dots+a_{n-1})} & \dots & 0 \end{bmatrix}.$$

We begin by transforming C so that its eigenvalues can be identified. First, let $D = \text{diag}(1, e^{ia_1}, e^{i(a_1+a_2)}, \dots, e^{i(a_1+\dots+a_{n-1})})$. Then, $C_1 = DC D^{-1}$ has 0 on the diagonal, $+1$ everywhere above the diagonal and -1 everywhere below the diagonal. Since C and C_1 are similar, they have the same eigenvalues. Now, $C_1 = f(T)$ where $f(x) = (x^n - x)/(x - 1)$ and $T = (t_{i,j})$ with $t_{i,i+1} = 1$ ($i = 1, \dots, n-1$), $t_{n,1} = -1$, and all other entries zero. Since $T^n = -I$, the eigenvalues of any $f(T)$ are $f(\eta_1), f(\eta_2), \dots, f(\eta_n)$, where the η_j are the roots of $z^n = -1$. Such matrices are known as *skew-circulants* (see P. J. Davis, *Circulant Matrices*, Wiley, 1979, sect. 3.2.1, pp. 83–84).

It follows that the eigenvalues of C are of the form $i \cot((2k-1)\pi/2n)$ for $k = 1, 2, \dots, n$, and the eigenvalue of largest absolute value arises from $k = 1$, corresponding to the largest eigenvalue of H . The Rayleigh-Ritz theorem (see Roger A. Horn & Charles R. Johnson, *Matrix Analysis*, Cambridge, 1985, theorem 4.2.2, p. 176) gives $\cot(\pi/2n) = \sup_{\mathbf{z} \neq \mathbf{0}} \langle H\mathbf{z}, \mathbf{z} \rangle / \langle \mathbf{z}, \mathbf{z} \rangle$. Separate real and imaginary parts as $H = A + iB$ and $\mathbf{z} = \mathbf{x} + i\mathbf{y}$. Then, Restriction to vectors with $\mathbf{y} = \mathbf{0}$ shows that this is an upper bound on the largest eigenvalue of A . As in solution I, this gives the desired bound on $\|A\|$.

No other solutions were received.

Coupled Squares

10213 [1992, 361]. *Proposed by P.G. Walsh, University of Waterloo, Waterloo, Ontario, Canada.*

Suppose x and y are positive integers such that $x + xy$ and $y + xy$ are both squares.

(a) Prove that exactly one of x and y is a square.

(b) Characterize all such pairs of integers x, y .

Solution by Robin J. Chapman, University of Exeter, Exeter, U. K. Let \mathbb{N} denote the set of positive integers.

(a) We first note that x and y cannot both be squares. For if x is a square, then $y + 1$ is also a square (since $x(y + 1)$ is a square), making it impossible, as $y \in \mathbb{N}$, for y to be a square.

Suppose now that $x, y \in \mathbb{N}$ and $x(y + 1)$ and $y(x + 1)$ are both squares. Write $x = ac^2$ and $x + 1 = bd^2$ where a and b are both squarefree. Then $ac^2(y + 1)$ and bd^2y are both squares, and so $y + 1 = au^2$ and $y = bv^2$ for some $u, v \in \mathbb{N}$. Hence

$$1 = au^2 - bv^2 \text{ and } -1 = ac^2 - bd^2. \quad (1)$$

It follows that $(auc + bvd)^2 - ab(cv + ud)^2 = (au^2 - bv^2)(ac^2 - bd^2) = -1$ and so the negative Pell equation $X^2 - abY^2 = -1$ has integer solutions. Note that ab is squarefree as a and b are coprime, and that $ab \neq 1$ as $x(x + 1)$ is not a square. Now by the theory of Pell's equation, if $r^2 - abs^2 = 1$ with $r, s \in \mathbb{N}$ then $(r + s\sqrt{ab}) = (p + q\sqrt{ab})^2$ where p and q are integers with $p^2 - abq^2 = \pm 1$. Apply this with $r = 2x + 1$ and $s = 2cd$. Writing $(r + s\sqrt{ab}) = (p + q\sqrt{ab})^2$ with $p^2 - abq^2 = \pm 1$ gives $2x + 1 = p^2 + abq^2$. Combining these two conditions on p and q , it follows that either $p^2 = x + 1$ or $p^2 = x$. In the latter case x is a square. In the former, $x + 1$ is a square and as $y(x + 1)$ is square, it follows that y is a square.

(b) Suppose by symmetry that $x = s^2$ is a square. Then $y + 1 = u^2$ is also a square. Now $(x + 1)y = (s^2 + 1)(u^2 - 1)$ is a square and if k is the greatest common divisor of $s^2 + 1$ and $u^2 - 1$ then

$$s^2 + 1 = kt^2 \text{ and } u^2 - 1 = kv^2 \quad (2)$$

for some $t, v \in \mathbb{N}$. Thus, s and t give a solution of the negative Pell equation $s^2 - kt^2 = -1$. By standard theory there are either zero or infinitely many such solutions for each $k \in \mathbb{N}$. Conversely, given such a k , one can find positive integers s, t, u and v satisfying (2). Then $x = s^2$ and $y = kv^2$ will have the required property.

Editorial comment. Dennis R. Estes observed that (1) gives rise to an equivalence between six ambiguous forms of discriminant ab . The fact that $a = 1$ or $b = 1$ follows from the work of Gauss showing that there are at most four ambiguous forms in the same class. Connections with this theory will be found in any solution.

Several readers invoked the theory of Continued Fractions to give a more detailed characterization of those $k \in \mathbb{N}$ for which the negative Pell equation is solvable.

Richard Stong took a different approach, which we sketch here. Eliminate y from the equations $x + xy = m^2$ and $y + xy = n^2$ to obtain $x(1 + x) + xn^2 = (1 + x)m^2$. Fixing x in this equation and looking at solutions (m, n) , one finds that (m', n') is also a solution with $m' = (2x + 1)m - 2xn$ and $n' = (2x + 1)n - 2(x + 1)m$ (also using the theory of the Pell equation), and y' can be found so that $x + xy' = m'^2$ and $y' + xy' = n'^2$. An easy analysis shows that $0 < m' < m$ if $n > m > 0$. This reduction, combined with the operation of interchanging x and y (with the corresponding interchange of m and n) allows every solution to be linked to a degenerate solution in which $x = 0$. Retracing one's steps allows the general solutions to be created out of the degenerate solution. The required property is proved by showing that it is preserved by these steps.

Solved also by J. Anglesio (France), D. R. Estes, N. J. Fine, I. Kastanas, K. S. Kedlaya (student), O. P. Lossers (The Netherlands), L. E. Mattics, W. W. Meyer, J. P. Robertson & J. B. Robertson, R. Stong, M. Vowe (Switzerland), and the proposer. Six incorrect or incomplete solutions were received.

A Maximum on the Boundary

10222 [1992, 462]. *Proposed by Gerry Myerson, Macquarie University, North Ryde, NSW, Australia.*

(a) Let h be a strictly increasing convex function on $[0, 1]$. Let n be a positive integer.

Assume that $0 \leq a_1 \leq \dots \leq a_n \leq 1$ and $0 \leq x_1 \leq \dots \leq x_n \leq 1$. Prove that

$$\sum_{j=1}^n h(|x_j - a_j|) \leq \max \left(\sum_{j=1}^n h(a_j), \sum_{j=1}^n h(1 - a_j) \right).$$

(b) Let n be a positive integer and let $a_j = (2j - 1)/2n$ for $1 \leq j \leq n$. Assume that $0 \leq x_1 \leq \dots \leq x_n \leq 1$. Let h be a strictly increasing, but not necessarily convex, function on $[0, 1]$. Prove that

$$\sum_{j=1}^n h(|x_j - a_j|) \leq \sum_{j=1}^n h(a_j).$$

Solution by Robin J. Chapman, University of Exeter, Exeter, U. K. (a) We use induction on n . If $n = 1$ then either $x_1 \leq a_1$ and so $h(|x_1 - a_1|) = h(a_1 - x_1) \leq h(a_1)$, or $x_1 \geq a_1$ and so $h(|x_1 - a_1|) = h(x_1 - a_1) \leq h(1 - a_1)$. Assume now that $n > 1$.

If $x_1 \leq a_1$ then $h(|x_1 - a_1|) = h(a_1 - x_1) \leq h(a_1)$ and the inductive hypothesis gives either $\sum_{j=1}^n h(|x_j - a_j|) \leq \sum_{j=1}^n h(a_j)$ or $\sum_{j=1}^n h(|x_j - a_j|) \leq h(a_1) + \sum_{j=2}^n h(1 - a_j)$. In the latter case if $a_1 \leq 1/2$ then $a_1 \leq 1 - a_1$ implying that $h(a_1) \leq h(1 - a_1)$ and so the result follows. On the other hand if $a_1 > 1/2$ then for $j > 1$, $a_j \geq a_1 > 1/2$ and $1 - a_j < a_j$. Hence $h(a_1) + \sum_{j=2}^n h(1 - a_j) \leq \sum_{j=1}^n h(a_j)$ and again the result follows.

A similar argument works if $x_n \geq a_n$.

Now if $x_1 > a_1$ and $x_n < a_n$ then there exists k with $a_k \leq x_k \leq x_{k+1} \leq a_{k+1}$. Hence $h(|x_k - a_k|) + h(|x_{k+1} - a_{k+1}|) = h(x_k - a_k) + h(a_{k+1} - x_{k+1})$. Now as h is convex then if $b, c > 0$ and $b + c \leq 1$, then

$$h(b) + h(c) \leq \frac{ch(0) + bh(b+c)}{b+c} + \frac{bh(0) + ch(b+c)}{b+c} = h(0) + h(b+c).$$

It follows that

$$\begin{aligned} h(|x_k - a_k|) + h(|x_{k+1} - a_{k+1}|) &\leq h(0) + h(a_{k+1} - a_k) \\ &\leq \min(h(a_k) + h(a_{k+1}), h(1 - a_k) + h(1 - a_{k+1})) \end{aligned}$$

If $n = 2$, we are done; else, by induction

$$\sum_{j \neq k, k+1}^n h(|x_j - a_j|) \leq \max \left(\sum_{j \neq k, k+1} h(a_j), \sum_{j \neq k, k+1} h(1 - a_j) \right),$$

and the result is now immediate.

(b) Let $I_j = [(j-1)/n, j/n]$ for $1 \leq j < n$ and $I_n = [(n-1)/n, 1]$. Then if $x_j \in I_{r_j}$ for each j , then $1 \leq r_1 \leq r_2 \leq \dots \leq r_n \leq n$. Also if $s_j = |r_j - j| + 1$ then $|x_j - a_j| \leq (2s_j - 1)/2n = a_{s_j}$. I claim that there is a permutation $\pi \in S_n$ such that $s_j \leq \pi(j)$ for all j . Given this claim it follows that

$$\sum_{j=1}^n h(|x_j - a_j|) \leq \sum_{j=1}^n h(a_{s_j}) \leq \sum_{j=1}^n h(a_{\pi(j)}) = \sum_{j=1}^n h(a_j),$$

as required.

It only remains to prove the claim. We again use induction on n . If $n = 1$ the claim is trivial. Suppose then that $n > 1$. We first suppose that $r_n < n$. Then if $j \leq n-1$ then $r_j \leq n-1$ and by induction there exists $\pi \in S_{n-1}$ with $s_j \leq \pi(j)$ for $1 \leq j \leq n-1$. Also $s_n \leq |1 - n| + 1 = n$ and so putting $\pi(n) = n$ we get the required permutation. Now suppose that $r_n = n$. If $j < n$ put $r'_j = r_j$ if $r_j < n$ and $r'_j = n-1$ if $r_j = n$. As $1 \leq r'_1 \leq r'_2 \leq \dots \leq r'_{n-1} \leq n-1$ then by induction there is a permutation $\rho \in S_{n-1}$

with $|r'_j - j| + 1 \leq \rho(j)$ for $1 \leq j < n$. Hence $s_j = |r_j - j| + 1 \leq \rho(j) + 1$. Now $s_n = |n - n| + 1 = 1$ and so if we put $\pi(j) = \rho(j) + 1$ for $j < n$ and $\pi(n) = 1$ then π is the required permutation.

Note that in neither (a) nor (b) do we need strict monotonicity of h .

Editorial comment. The proposer used the assumption of strict inequality to guarantee that the sum would have a maximum attained at a single point. Most solvers succumbed to the temptation to simplify the argument by characterizing the point (x_1, \dots, x_n) at which the maximum is attained.

Solved also by M. V. Bjelica (Yugoslavia), M. Bowron, R. B. Israel (Canada), M. Mócsy (Hungary), K. Schilling, and the proposer.

An Integral Infinite Sum

10231 [1992, 570]. *Proposed by Adrian Riskin, Northern Arizona University, Flagstaff, AZ.*

For positive integers m and n , let

$$f(m, n) = \sum_{k=1}^{\infty} k^n \left(\frac{m}{m+1}\right)^k.$$

(a) Prove that $f(m, n)$ is an integer.

(b) Show that the last digit of the decimal expansion of $f(1, n)$ can only be 0, 2, or 6.

Solution to part (a) by David Beckwith, Sag Harbor, NY. Let

$$g_n(x) = f(x, n) = \sum_{k=1}^{\infty} k^n \left(\frac{x}{x+1}\right)^k.$$

The series converges uniformly on a positive interval; termwise differentiation yields

$$g'_n(x) = \sum_{k=1}^{\infty} k^n k \left(\frac{x}{x+1}\right)^{k-1} \frac{1}{(x+1)^2} = \frac{g_{n+1}(x)}{x(x+1)}.$$

Hence $g_{n+1}(x) = x(x+1)g'_n(x)$. By explicit computation, $g_0(x) = x$. By the recurrence, every $g_n(x)$ is a polynomial in x with integer coefficients. Hence every $f(m, n) = g_n(m)$ is an integer.

Solution to part (b) by Richard Holzsager, American University, Washington, DC. Starting with $f(m, 1) = g_1(m) = m^2 + m$ and applying the recurrence, we obtain the additional polynomials

$$f(m, 2) = 2m^3 + 3m^2 + m,$$

$$f(m, 3) = 6m^4 + 12m^3 + 7m^2 + m,$$

$$f(m, 4) = 24m^5 + 60m^4 + 50m^3 + 15m^2 + m,$$

$$f(m, 5) = 120m^6 + 360m^5 + 390m^4 + 180m^3 + 31m^2 + m.$$

Reducing the coefficients modulo 10, we have $f(m, 5) \equiv f(m, 1) \pmod{10}$. Hence the succeeding polynomials repeat mod 10 with a period of 4, for any fixed m . For $m = 1$, the cycle is 2, 6, 6, 0.

Editorial comment. A popular method of solution was to show that

$$f(m, n) = (m+1) \sum_{k=1}^n k! S(n, k) m^k,$$

where $S(n, k)$ denotes the Stirling numbers of the second kind. This establishes (a), and $f(1, n) \equiv 2 \sum_{k=1}^4 k! S(n, k) \pmod{10}$ leads to (b).

Gerry Meyerson noted that $f(m, n)$ is an integer multiple of $m(m+1)$ for $n > 0$. This also follows from the selected proof. He also located $f(1, n)/2$ as sequence #1191 in N. J. A. Sloane, *Handbook of Integer Sequences*, Academic Press, 1973, where the sequence is traced back to Cayley. The number $f(1, n)/2$ is equal to the number of distributions of n distinct objects into ordered cells such that no occupied cell is above an unoccupied cell. A proof of part (b) using this interpretation can be found in O. A. Gross, "Preferential arrangements", this MONTHLY 69 (1962), 4-8. This latter reference was mentioned by István Nemes.

William Y. C. Chen gave further references dealing with the question of periodicity of the $f(1, n)$ modulo primes.

Solved by 54 solvers and the proposer.

Mutually Convergent Series

10291 [1993, 290]. *Proposed by Howard Morris, Chatsworth, CA.*

Let k be a positive integer and let $\langle x_n \rangle$ be a nondecreasing sequence of real numbers for which $\sum (1/x_n)$ converges. Show that $\sum (\ln x_n)^k / x_n$ converges if and only if $\sum (\ln n)^k / x_n$ converges.

Solution by Frank Schmidt, Arlington, VA. Since $x_n \leq x_{n+1}$ and $\sum (1/x_n) < \infty$, we have $\lim_{n \rightarrow \infty} (n/x_n) = 0$ (see *Editorial Comment* below). In particular, there is a real number K with $n \leq Kx_n$ for all n , so that $\sum (\log x_n)^k / x_n < \infty$ implies $\sum (\log n)^k / x_n < \infty$. To prove the converse, split the indices into two subsets: (I) $x_n \leq n^{k+2}$; (II) $x_n > n^{k+2}$. On subset I, $\log x_n \leq (k+2)(\log n)$, hence $\sum_I (\log n)^k / x_n < \infty$ implies $\sum_I (\log x_n)^k / x_n < \infty$. As for subset (II), for sufficiently large n (depending on k), we have

$$\frac{(\log x_n)^k}{x_n} < \frac{(x_n)^{\frac{k}{k+1}}}{x_n} = \frac{1}{(x_n)^{\frac{1}{k+1}}} < \frac{1}{n^{\frac{k+2}{k+1}}}.$$

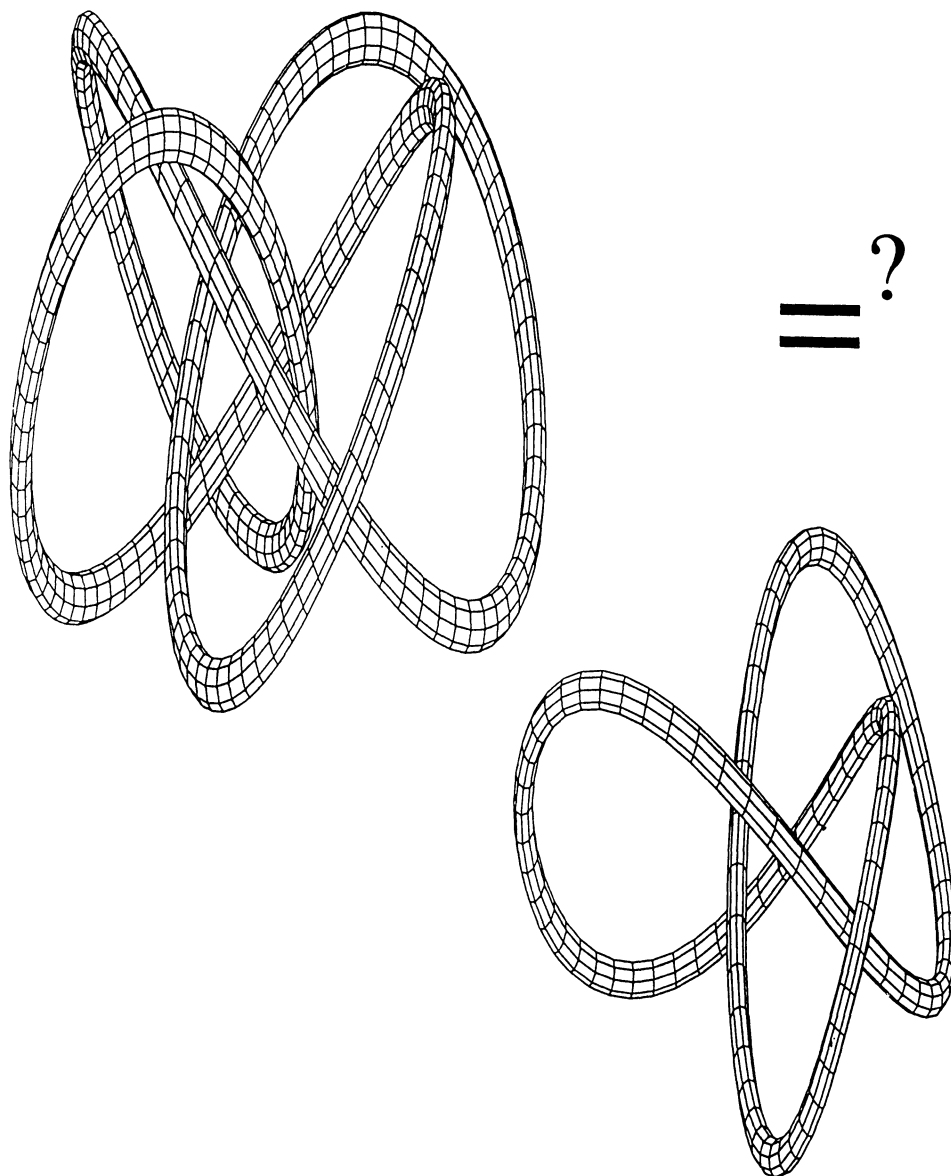
Since $(k+2)/(k+1) > 1$, $\sum_{II} (\log x_n)^k / x_n < \infty$.

Editorial comment. The upper bound on x_n may be obtained in many ways: most readers gave a simple *ad hoc* proof; three readers appeared to treat it as *obvious*; and three readers referred to a well-known result (theorems of Abel, Kroneker or Pringsheim) without citing the statement that the solver had in mind. An editor supplied another approach: paraphrasing theorem 3.27 (attributed to Cauchy) of Walter Rudin, *Principles of Mathematical Analysis*, 3rd edition, McGraw-Hill, 1976, we find that given conditions imply $\sum 2^k / x_{2^k}$ converges, from which the result follows.

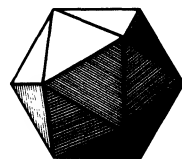
H.-J. Seiffert observed that the upper bound on x_n is also part of the general theory of convergence exponent λ of $\langle x_n \rangle$. For x_n as in this problem, λ is characterized by $\sum x_n^{-\sigma}$ converging for $\sigma > \lambda$ and diverging for $\sigma < \lambda$, so that $\lambda \leq 1$ in this case. In G. Pólya and G. Szegő, *Problems and Theorems in Analysis*, Vol. II, Springer-Verlag, 1972-76, pp. 25-26, entry 113, one finds that $\lambda = \limsup_{n \rightarrow \infty} \frac{\log n}{\log x_n}$. Although this leads to slightly weaker inequalities than have been mentioned above, it is strong enough for the present needs.

Solved also by V. Božin (student, Yugoslavia), D. A. Darling, E. Hertz, R. Holzager, G. L. Isaacs, I. Kastanas, A. D. Melas (Greece), A. Pedersen (Denmark), H.-J. Seiffert (Germany), R. Stong, A. A. Tarabay (Lebanon), R. B. Tucker, H. V. Vu (student, Hungary), A. N. 't Woord (the Netherlands), and the proposer.

Collaborating editors: *David F. Appleyard, Paul T. Bateman, Bruce C. Berndt, Duane M. Broline, Barry W. Brunson, Frank S. Cater, Gulbank D. Chakerian, Underwood Dudley, Gerald A. Edgar, Michael A. Filaseta, Ira M. Gessel, Richard A. Gibbs, Jerrold R. Griggs, Douglas A. Hensley, John R. Isbell, Mourad E. H. Ismail, Murray Klamkin, Daniel J. Kleitman, Frederick W. Luttman, Frank B. Miles, Richard Pfeifer, Stephen L. Portnoy, J. O. Shallit, John Henry Steelman, Kenneth B. Stolarsky, David E. Tepper, Douglas B. Tyler, Daniel Ullman, and William E. Watkins.*



The American Mathematical Monthly



Volume 102, Number 3 / MARCH 1995



8
5 3
5 6
2 9
9 5
1 4
4 1
3

A Spigot Algorithm for π
(see page 195)

NOTICE TO AUTHORS

The *Monthly* publishes articles, notes, and other features about mathematics and the profession. The readership of the *Monthly* is intended to include everybody who is mathematically inclined, including of course professional mathematicians and students of mathematics at all collegiate levels. While no single article or feature is likely to appeal to everyone, material should interest and be accessible to a large number of readers. This is the most important criterion for acceptance.

Articles may be expositions of old results or presentations of new ones. They may concern all of mathematics or one small area, a broad development or a single application, historical reminiscences or one important event. While some articles may contain the author's new research, the novelty of material and generality of the results is far less important than the clarity of exposition and general interest. Discussing one illuminating case of a well known result is far better than providing all the details of an obscure but new proposition. Articles in the *Monthly* are supposed to inform and to entertain; they are meant to be read rather than archived.

Notes are short and possibly informal articles. A note may concern a clever new proof of an old theorem, a novel way to present tired material, or a lively discussion of a philosophical (but still mathematical) issue. Also, any topic is suitable, so long as it is related to mathematics. Because a note is short, the first few sentences are the most important part: They should explain the purpose and invite the reader in. Photographs or diagrams often will attract the reader's attention.

All articles and notes should be sent to the editor:

JOHN EWING
Department of Mathematics
Indiana University
Bloomington, IN 47405

Please send 3 copies, typewritten on only one side of the paper. Illustrations should be carefully drawn on separate sheets of paper in black ink; the original should be without lettering and two copies should have appropriate captions and lettering indicated.

Proposed problems or solutions should be sent to:

RICHARD BUMBY,
P.O. Box 10971
New Brunswick, NJ 08906-0971.

Please send 2 copies of all material, typewritten if possible.

Letters to the Editor, both for publication and for private reading, should be sent to the Editor at the address given above. Comments, including criticisms, are welcome, as are all suggestions for making the *Monthly* a lively, entertaining, and informative journal.

EDITOR:

JOHN H. EWING

ASSOCIATE EDITORS:

PETER BORWEIN	FRED KOCHMAN
RICHARD BUMBY	CATHERINE MCGEOCH
DENNIS DETURCK	RICHARD NOWAKOWSKI
UNDERWOOD DUDLEY	ARNOLD OSTEBEE
JOHN DUNCAN	LEE RUBEL
JOAN FERRINI-MUNDY	ABE SHENITZER
JOSEPH GALLIAN	LYNN STEEN
STEVEN GALOVICH	STAN WAGON
RICHARD GUY	DOUGLAS WEST
DARRELL HAILE	HERBERT WILF
PAUL HALMOS	SANDY ZABELL
JOAN HUTCHINSON	PAUL ZORN

EDITORIAL ASSISTANT:

MISTY CUMMINGS

STAFF ARTIST:

MIKE CAGLE

QUOTE MASTER:

MARK WOODARD

Reprint permission:

MARCIA P. SWARD, Executive Director

Advertising Correspondence:

Ms. ELAINE PEDREIRA, Advertising Manager

Subscription correspondence, change of address, and other inquiries:

Membership / Subscriptions Department

All at the address:

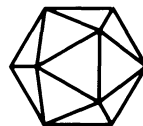
The Mathematical Association of America
1529 Eighteenth Street, N.W.
Washington, DC 20036.

Microfilm Editions: University Microfilms International, Serial Bid coordinator, 300 North Zeeb Road, Ann Arbor, MI 48106.

The AMERICAN MATHEMATICAL MONTHLY (ISSN 0002-9890) is published monthly except bimonthly June-July and August-September by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, DC 20036 and Montpelier, VT. Copyrighted by the Mathematical Association of America (Incorporated), 1995, including rights to this journal issue as a whole and, except where otherwise noted, rights to each individual contribution. General permission is granted to Institutional Members of the MAA for noncommercial reproduction in limited quantities of individual articles (in whole or in part) provided a complete reference is made to the source. Second class postage paid at Washington, DC, and additional mailing offices. **Postmaster:** Send address changes to the American Mathematical Monthly, Membership / Subscription Department, MAA, 1529 Eighteenth Street, N.W., Washington, DC, 20036-1385.

**The American
Mathematical Monthly**

Volume 102, Number 3 / MARCH 1995
(ISSN 0002-9890)



Contents

ARTICLES

A Spigot Algorithm for the Digits of π / STANLEY RABINOWITZ and STAN WAGON 195

The Rise, Fall, and Possible Transfiguration of Triangle Geometry:
A Mini-history / PHILIP J. DAVIS 204

Totally Real Origami and Impossible Paper Folding / DAVID AUCKLY
and JOHN CLEVELAND 215

An Abstract Algebra Story / URI LERON and ED DUBINSKY 227

A Multidimensional Version of Rolle's Theorem / MASSIMO FURI and MARIO MARTELLI 243

FEATURES

COMMENTS 194

NOTES

Adding Distinct Congruence Classes Modulo a Prime / NOGA ALON,
MELVYN B. NATHANSON, and IMRE RUZSA 250

A Simple Proof of the Hölder and the Minkowski Inequality /
LECH MALIGRANDA 256

THE COMPUTER SCIENCE SAMPLER

Missing Real Numbers / CHRISTOPHER J. VAN WYK 260

THE EVOLUTION OF ...

The Evolution of Algebra 1800–1870 / I. G. BASHMAKOVA
and A. N. RUDAKOV 266

THE AUTHORS 271

PROBLEMS AND SOLUTIONS 273

REVIEWS

Squares. By A. R. Rajwade / DANIEL B. SHAPIRO 281

TELEGRAPHIC REVIEWS 285

A Spigot Algorithm for the Digits of π

Stanley Rabinowitz and Stan Wagon

It is remarkable that the algorithm illustrated in Table 1, which uses no floating-point arithmetic, produces the digits of π . The algorithm starts with some 2s, in columns headed by the fractions shown. Each entry is multiplied by 10. Then, starting from the right, the entries are reduced modulo *den*, where the head of the column is *num/den*, producing a quotient *q* and remainder *r*. The remainder is left in place and $q \times \text{num}$ is carried one column left. This reduce-and-carry is continued all the way left. The tens digit of the leftmost result is the next digit of π . The process continues with the multiplication of the remainders by 10, the reductions modulo the denominators, and the augmented carrying.

TABLE 1. The workings of an algorithm that produces digits of π . The dashed line indicates the key step: starting from the right, entries are reduced modulo the denominator of the column head (25, 23, 21, ..., resp.), with the quotients, after multiplication by the numerator (12, 11, 10, ...), carried left. For example, the 20 in the $\frac{9}{19}$'s column yields a remainder of 1 and a left carry of $1 \cdot 9 = 9$. After the leftmost carries, the tens digits are 3, 1, 4, 1. To get more digits of π one must start with a longer string of 2s.

	Digits of π	$\frac{1}{3}$	$\frac{2}{5}$	$\frac{3}{7}$	$\frac{4}{9}$	$\frac{5}{11}$	$\frac{6}{13}$	$\frac{7}{15}$	$\frac{8}{17}$	$\frac{9}{19}$	$\frac{10}{21}$	$\frac{11}{23}$	$\frac{12}{25}$
Initialize		2	2	2	2	2	2	2	2	2	2	2	2
$\times 10$		20	20	20	20	20	20	20	20	20	20	20	20
Carry	3	<u>10</u>	<u>12</u>	<u>12</u>	<u>12</u>	<u>10</u>	<u>12</u>	<u>7</u>	<u>8</u>	<u>9</u>	<u>0</u>	<u>0</u>	<u>0</u>
Remainders		30	32	32	32	30	32	27	28	29	20	20	20
		0	2	2	4	3	10	1	13	12	1	20	20
$\times 10$		0	20	20	40	30	100	10	130	120	10	200	200
Carry	1	<u>13</u>	<u>20</u>	<u>33</u>	<u>40</u>	<u>65</u>	<u>48</u>	<u>98</u>	<u>88</u>	<u>72</u>	<u>150</u>	<u>132</u>	<u>96</u>
Remainders		13	40	53	80	95	148	108	218	192	160	332	296
		3	1	3	3	5	5	4	8	5	8	17	20
$\times 10$		30	10	30	30	50	50	40	80	50	80	170	200
Carry	4	<u>11</u>	<u>24</u>	<u>30</u>	<u>40</u>	<u>40</u>	<u>42</u>	<u>63</u>	<u>64</u>	<u>90</u>	<u>120</u>	<u>88</u>	<u>0</u>
Remainders		41	34	60	70	90	92	103	144	140	200	258	200
		1	1	0	0	0	4	12	9	4	10	6	16
$\times 10$		10	10	0	0	0	40	120	90	40	100	60	160
Carry	1	<u>4</u>	<u>2</u>	<u>9</u>	<u>24</u>	<u>55</u>	<u>84</u>	<u>63</u>	<u>48</u>	<u>72</u>	<u>60</u>	<u>66</u>	<u>0</u>
		14	12	9	24	55	124	183	138	112	160	126	160

This algorithm is a “spigot” algorithm: it pumps out digits one at a time and does not use the digits after they are computed. Moreover, the digits are generated without any use of high-precision (or low-precision) operations on floating-point real numbers; the entire algorithm uses only ordinary integer arithmetic on

relatively small integers. For example, to obtain the first 5,000 digits of π requires only arithmetic operations on integers less than 600,000,000. Although high-precision floating-point routines are built up from integer operations, the algorithms in this paper are quite simple and do not simulate floating-point computations.

In order to motivate the π -algorithm, we first discuss the much simpler case of e , for which a spigot algorithm was discovered by Sale [Sale]. His algorithm is the basis of the discussion in §1.

1. A NUMBER SYSTEM IN WHICH e 's DIGITS ARE PERIODIC. A real number's decimal representation may be interpreted as an infinitely nested expression; for example:

$$\sqrt{2} = 1.41421356\dots = 1 + \frac{1}{10}\left(4 + \frac{1}{10}\left(1 + \frac{1}{10}\left(4 + \frac{1}{10}\left(2 + \frac{1}{10}(1 + \dots)\right)\right)\right)\right).$$

Some interesting and useful representations may be obtained if we change the base-sequence, which in the case above is $(\frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \dots)$. For example, using the base $\mathbf{b} = (\frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \dots)$ yields the following form, called a *mixed-radix* representation (see [Knu, §4.1]):

$$a_0 + \frac{1}{2}\left(a_1 + \frac{1}{3}\left(a_2 + \frac{1}{4}\left(a_3 + \frac{1}{5}\left(a_4 + \frac{1}{6}(a_5 + \dots)\right)\right)\right)\right),$$

where the a_i (the *digits*) are nonnegative integers. If $0 \leq a_i \leq i$ for $i \geq 1$, the representation is called *regular*. Mixed-radix representations will be denoted by $(a_0; a_1, a_2, a_3, a_4, \dots)_{\mathbf{b}}$. For base \mathbf{b} , every positive real number has a regular representation and representations are unique provided we exclude representations that terminate with maximal digits (otherwise, for example, $\frac{1}{2} = (0; 1, 0, 0, \dots)_{\mathbf{b}} = (0; 0, 2, 3, 4, 5, 6, \dots)_{\mathbf{b}}$); from now on and for all bases, we exclude such representations. The proof of the following Lemma is in Appendix 1.

Lemma 1(a). *If $i \geq 1$, $(0; 0, 0, \dots, 0, a_i, a_{i+1}, \dots)_{\mathbf{b}} < \frac{1}{i!}$; in particular, $(0; a_1, a_2, a_3, a_4, \dots)_{\mathbf{b}} < 1$.*

(b). *Representations using the mixed-radix base \mathbf{b} are unique.*

(c). *The integer part of $(a_0; a_1, a_2, a_3, a_4, \dots)_{\mathbf{b}}$ is a_0 and the fractional part is $(0; a_1, a_2, a_3, a_4, \dots)_{\mathbf{b}}$.*

In this number system some irrationals become periodic. For example, $e = (2; 1, 1, 1, 1, \dots)_{\mathbf{b}}$; this is just a restatement of the infinite series $\sum \frac{1}{i!}$ as $1 + \frac{1}{1}(1 + \frac{1}{2}(1 + \frac{1}{3}(1 + \frac{1}{4}(1 + \frac{1}{5}(1 + \dots))))$). Rational numbers in this system correspond to digit-sequences that terminate (Appendix 1, Lemma 2).

The decimal digits of a real number x in $[0, 10)$ can be obtained by taking the integer part of x , multiplying its fractional part by 10, taking the integer part of the result, multiply the resulting fractional part by 10, and so on. In some mixed-radix bases, this is especially simple. If $x = (a_0; a_1, a_2, \dots, a_n)_{\mathbf{b}}$, then $10x = (10a_0; 10a_1, 10a_2, 10a_3, \dots, 10a_n)_{\mathbf{b}}$. The latter may not be a regular expression: some digits may be too big. But we can decrease digits by reducing them modulo i , where i is the denominator of the corresponding element of \mathbf{b} . Starting these reductions at the right end, we carry the quotients left, eventually getting the regular representation of $10x$. Thus multiplying by 10 is algorithmically straightforward. Taking the integer and fractional parts for \mathbf{b} -representations is also easy, thanks to Lemma 1(c).

We can now give the algorithm to get the first n base-10 digits of e . A proof of correctness—the error analysis showing that $n + 2$ mixed-radix digits suffice¹ to get n base-10 digits—is given as Lemma 3 in Appendix 1.

Algorithm e -spigot

1. *Initialize:* Let the first digit be 2 and initialize an array A of length $n + 1$ to $(1, 1, 1, \dots, 1)$.
2. Repeat $n - 1$ times:
Multiply by 10: Multiply each entry of A by 10.
Take the fractional part: Starting from the right, reduce the i th entry of A modulo $i + 1$, carrying the quotient one place left.
Output the next digit: The final quotient is the next digit of e .

The first few steps of this algorithm, starting with an array of 10 1s (this corresponds to 11 mixed-radix digits, good for 9 digits of e ; only 5 are shown), are displayed in Table 2.

TABLE 2. The workings of a spigot algorithm for the digits of e (in bold). The reductions in the column headed $\frac{1}{i}$ are performed modulo i . The leftmost base-10 real numbers are the values of the rows viewed as mixed-radix representations. Since only 11 mixed-radix digits start the algorithm, the first base-10 number is only an approximation to e .

Base 10		$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{7}$	$\frac{1}{8}$	$\frac{1}{9}$	$\frac{1}{10}$	$\frac{1}{11}$
2.718281826...	2	1	1	1	1	1	1	1	1	1	1
7.18281826...		10	10	10	10	10	10	10	10	10	10
carries	7	± 3	± 3	± 2	± 1	± 1	± 1	± 1	± 1	± 0	\pm
0.18281826...		14	13	12	11	11	11	11	11	10	10
1.8281826...		0	1	0	1	5	4	3	2	0	10
carries	1	± 3	± 0	± 3	± 9	± 6	± 4	± 2	± 0	± 9	\pm
0.8281826...		3	10	3	19	56	44	32	20	9	100
8.281826...		1	1	3	4	2	2	0	2	9	1
carries	8	± 6	± 9	± 8	± 3	± 2	± 0	± 3	± 9	± 0	\pm
0.281826...		16	19	38	43	22	20	3	29	90	10
2.81826...		0	1	2	3	4	6	3	2	0	10
carries	2	± 5	± 6	± 7	± 8	± 9	± 4	± 2	± 0	± 9	\pm
0.81826...		5	16	27	38	49	64	32	20	9	100
		1	1	3	3	1	1	0	2	9	1

2. A SPIGOT FOR DIGITS OF π . The ideas of §1 lead to a spigot algorithm for π , but there are additional complexities and additional interesting questions that distinguish π from e . Our starting point is the following moderately well-known

¹Any digit-producing algorithm for a presumed-normal number x suffers from a drawback that, although unlikely, can impinge on the result. If x is between 1 and 10 and the algorithm says that the first 100 digits of x are, say, 4, 6, 5, 0, 7, . . . , 3, 9, 9, 9, 9 then one cannot be sure that the last 6 digits are correct. They will be the digits of a certain approximation to x that is within $5 \cdot 10^{-100}$ of the true value. One cannot simply go farther until a non-9 is reached, because memory allocations must be made in advance. The user must realize that a terminating string of 9s is a red flag concerning those digits and even with no 9s, the last digit might be incorrect. In practice, one might ask for, say, 6 extra digits, reducing the odds of this problem to one in a million.

series:

$$\pi = \sum_{i=0}^{\infty} \frac{(i!)^2 2^{i+1}}{(2i+1)!}.$$

This series can be derived from the Wallis product for π ; another approach uses an acceleration technique called Euler's transform applied to the series $\pi = 4 - \frac{4}{3} + \frac{4}{5} - \frac{4}{7} + \dots$. These proofs, together with three others and references to earlier sources, may be found in [Li]. We let $k!!$ denote the product $1 \cdot 3 \cdot 5 \cdots k$ for odd integers k ; then the series is equivalent to

$$\frac{\pi}{2} = \sum_{i=0}^{\infty} \frac{i!}{(2i+1)!!} = 1 + \frac{1}{3} + \frac{1 \cdot 2}{3 \cdot 5} + \frac{1 \cdot 2 \cdot 3}{3 \cdot 5 \cdot 7} + \dots,$$

which expands to become

$$\frac{\pi}{2} = 1 + \frac{1}{3} \left(1 + \frac{2}{5} \left(1 + \frac{3}{7} \left(1 + \frac{4}{9} (1 + \dots) \right) \right) \right).$$

This last expression leads to the mixed-radix base $\mathbf{c} = (\frac{1}{3}, \frac{2}{5}, \frac{3}{7}, \frac{4}{9}, \dots)$, with respect to which π is simply $(2; 2, 2, 2, 2, \dots)_{\mathbf{c}}$. For a regular representation in base \mathbf{c} , the digit in the i th place must lie in the interval $[0, 2i]$. Unfortunately, base \mathbf{c} is less accommodating than \mathbf{b} .

Lemma 4 (Proof in Appendix 1). *The base- \mathbf{c} number with maximal digits, $(0; 2, 4, 6, 8, \dots)_{\mathbf{c}}$, represents 2; hence regular representations of the form $(0; a, b, c, \dots)_{\mathbf{c}}$ lie between 0 and 2.*

Lemma 4 implies that \mathbf{c} -representations are not unique. For example, $(0; 0, 4, 6, 8, \dots)_{\mathbf{c}} = 2 - \frac{2}{3} = \frac{4}{3}$, whence $(0; 0, 2, 3, 4, \dots)_{\mathbf{c}} = \frac{2}{3} = (0; 2, 0, 0, \dots)_{\mathbf{c}}$. More relevant algorithmically, integer and fractional parts using \mathbf{c} are not straightforward, as they are for \mathbf{b} . The integer part of $(a_0; a_1, a_2, \dots)_{\mathbf{c}}$ is either a_0 or $a_0 + 1$ according as $(0; a_1, a_2, \dots)_{\mathbf{c}}$ is in $[0, 1)$ or $[1, 2)$. This problem is surmounted by leaving the units digit of a_0 in place during the next iteration and calling the tens digit of a_0 a *predigit*. The predigits must be temporarily held because occasionally (once every 20 iterations, roughly) the next predigit is a 10; this will happen when the carry, which is between 0 and 19, is greater than 10 and, simultaneously, the leftover units digit of a_0 is 9, which becomes 90 in the multiply-by-10 step. This event requires that the held number be increased by 1 before being released. Specific details of the algorithm follow; the presentation at the beginning of this paper sidestepped the problem of the occasional 10. The proof that $\lceil 10n/3 \rceil$ mixed-radix digits suffice for n digits of π is in Appendix 1 (Lemma 5). Appendix 2 contains a Pascal implementation of this algorithm.

Algorithm π -spigot

1. *Initialize:* Let $A = (2, 2, 2, 2, \dots, 2)$ be an array of length $\lceil 10n/3 \rceil$.
2. Repeat n times:
 - Multiply by 10:* Multiply each entry of A by 10.
 - Put A into regular form:* Starting from the right, reduce the i th element of A (corresponding to \mathbf{c} -entry $(i-1)/(2i-1)$) modulo $2i-1$, to get a quotient q and a remainder r . Leave r in place and carry $q(i-1)$ one place left. The last integer carried (from the position where $i-1=2$) may be as large as 19.

Get the next predigit: Reduce the leftmost entry of A (which is at most $109 [= 9 \cdot 10 + 19]$) modulo 10. The quotient, q , is the new predigit of π , the remainder staying in place.

Adjust the predigits: If q is neither 9 nor 10, release the held predigits as true digits of π and hold q . If q is 9, add q to the queue of held predigits. If q is 10 then:

- set the current predigit to 0 and hold it;
- increase all other held predigits by 1 (9 becomes 0);
- release as true digits of π all but the current held predigit.

This algorithm uses only integer arithmetic and is easy to program. The table at the beginning of the paper shows it in action, starting with 13 mixed-radix digits of π (good for 4 base-10 digits). To clarify the working of the algorithm, note that the (finite) first row of Table 1 is a mixed-radix representation of $3.1414796\dots$, the second row represents $31.414796\dots$, the fifth row represents $1.414796\dots$, the sixth row is $14.14796\dots$, the ninth row is $4.14796\dots$, and so on. Table 3 shows the result of a computation using a larger initial array; the holding aspect does not become relevant until the 32nd digit.

TABLE 3. The actual digits of π (bottom) compared to the sequence of leftmost base- c digits for 35 iterations with a starting array of 116 2s (good for 35 digits). At the 32nd iteration a 102 shows up, yielding a predigit of 10.

30	13	41	15	58	92	26	64	53	35	58	89	97	78	92	32	23	38	84	45	62	26	63	42	33	38	82	32	27	78	94	49	102	28	87
3	1	4	1	5	9	2	6	5	3	5	8	9	7	9	3	2	3	8	4	6	2	6	4	3	3	8	3	2	7	9	5	0	2	8

We repeat that the algorithm uses only integer operations. To get 5,000 digits of π requires only integer arithmetic on numbers less than 600,000,000. The algorithm leads naturally to the question of improving it to one that is essentially as simple as *e-spigot*.

Question. Is there a base d of rationals such that π has a d -representation that is periodic, or an arithmetic progression, and such that a_0 is always the integer part of $(a_0; a_1, a_2, \dots)_d$?

Gosper [Gos, p. 32] has discovered a series for π that brings us tantalizingly close to spigot-perfection:

$$\pi = 3 + \frac{1}{60}8 + \frac{1}{60} \frac{2 \cdot 3}{7 \cdot 8 \cdot 3}13 + \frac{1}{60} \frac{2 \cdot 3}{7 \cdot 8 \cdot 3} \frac{3 \cdot 5}{10 \cdot 11 \cdot 3}18 + \frac{1}{60} \frac{2 \cdot 3}{7 \cdot 8 \cdot 3} \frac{3 \cdot 5}{10 \cdot 11 \cdot 3} \frac{4 \cdot 7}{13 \cdot 14 \cdot 3}23 + \dots$$

He obtained this series by using a refinement of the Euler transform on $4 - \frac{4}{3} + \frac{4}{5} - \frac{4}{7} + \dots$. Gosper's series leads to the base $d = (\frac{1}{60}, \frac{6}{168}, \frac{15}{330}, \frac{28}{546}, \dots)$, with respect to which π is $(3; 8, 13, 18, \dots)$. A computation shows that $(0; 59, 167, 329, 545, \dots)_d = 1.092\dots$, a substantial improvement over the 2 that arose for c . Under the usual randomness assumption for π 's digits, the odds of a

bad predigit in base **c** are 1 in 20, while in base **d** they decrease to less than 1 in 110; this is because a **d**-predigit of 10 occurs only when the remainder is a 9 (which becomes 90) and the carry is a 10. The former happens 10% of the time, while the latter happens no more than once in 11 iterations because the carry is the integer part of a real between 0 and 10.93. So base **d** is within 1% of spigot-perfection. Because Gosper's series converges more quickly than the one we used, it has less memory requirements: n digits of π require an initial array of length n ; however, the arithmetic on the array will involve integers larger than those in an array of the same size using base **c**.

One way to improve the Gosper-series approach is to reduce the fractions in **d** to lowest terms. Then the regular number with maximal digits is $(0; 59, 27, 21, 38, \dots)_d$, which equals $1.0000476468\dots$. It is not hard to see that the regular representation of π is unchanged in this new base. However, the work expended in reducing to lowest terms outweighs the gain made in reducing the number of times a 10 appears as a predigit. Thus it is likely that an affirmative answer to the question above is of more theoretical than practical interest.

The spigot algorithm for π is by no means competitive with the recently discovered fast algorithms (due to the Borwein brothers, the Chudnovsky brothers, and others) that have been used to compute hundreds of millions of digits of π (see [BBB]). But the spigot algorithm does have the advantage of avoiding all floating-point computations; thus it is easily implemented on a home computer where it can produce thousands of digits in a few minutes. Moreover, it gives the result directly in base 10 (most other π -algorithms produce the result in binary or some internal format and a second pass must be made to obtain decimal digits).

The algorithm given here can be made to run faster by outputting multiple digits at a time. For example, to get five decimal digits at a time, simply compute the digits of π using base 100,000. This can be done by multiplying by 100,000 instead of 10 in the main step. The integer part is then the next "digit" in base 100,000.

If one is working in base 100,000 and knows in advance that the portion of digits to be computed does not contain the string 00000, then one can omit the lengthy part of the algorithm that adjusts the predigits. This can lead to an exceedingly short computer program. For example, Rabinowitz [Rab] used this idea to exhibit a 14-line Fortran program that outputs 1,000 decimal digits of π .

Finally, we mention that the algorithm can be parallelized, in which case it becomes blindingly fast up to about 10,000 digits.

For examples of spigot algorithms for other functions, see [Abd].

APPENDIX 1. FIVE LEMMAS

Lemma 1(a). *If $i \geq 1$, $(0; 0, 0, \dots, 0, a_i, a_{i+1}, \dots)_b < \frac{1}{i}$; in particular, $(0; a_1, a_2, a_3, a_4, \dots)_b < 1$.*

(b). *Representations using mixed-radix base **b** are unique.*

(c). *The integer part of $(a_0; a_1, a_2, a_3, a_4, \dots)_b$ is a_0 and the fractional part is $(0; a_1, a_2, a_3, a_4, \dots)_b$.*

Proof: (a). It suffices to prove that $\sum_{k=i+1}^{\infty} (k-1)/k! = 1/i!$, which follows from the fact that the series telescopes to:

$$\left(\frac{1}{i!} - \frac{1}{(i+1)!} \right) + \left(\frac{1}{(i+1)!} - \frac{1}{(i+2)!} \right) + \left(\frac{1}{(i+2)!} - \frac{1}{(i+3)!} \right) + \dots$$

(b). Suppose $(a_0; a_1, a_2, a_3, a_4, \dots)_b$ and $(c_0; c_1, c_2, c_3, c_4, \dots)_b$ represent the same real number. Then, for some i , $0 = \sum_{k=i}^{\infty} d_k/k!$, where $|d_k| < k$ and $d_i \neq 0$. But then $|d_i|/i! \leq \sum_{k=i+1}^{\infty} |d_k|/k!$, contradicting (a).

(c). This follows from (a).

Lemma 2. *A positive number is rational iff its digits using the mixed-radix base \mathbf{b} are eventually 0.*

Proof: The reverse direction is obvious. For the forward direction we use a sublemma.

Sublemma. *For any integers t and n , with $0 \leq n < t!$, there are integers d_i in $[0, i]$ such that $n = d_1 t(t-1)(t-2) \cdots 4 \cdot 3 + d_2 t(t-1)(t-2) \cdots 5 \cdot 4 + \cdots + d_{t-3} t(t-1) + d_{t-2} t + d_{t-1}$.*

Proof: By induction on t . If $n < t!$ write n as $qt + r$ with $0 \leq r < t$ and $0 \leq q < (t-1)!$. By induction there is a sequence $(d_1, d_2, \dots, d_{t-3}, d_{t-2})$ that is a solution for q with respect to terms $(t-1)(t-2) \cdots 4 \cdot 3$, and the like, whence $(d_1, d_2, \dots, d_{t-3}, d_{t-2}, r)$ is a solution for n w.r.t. the terms $t(t-1)(t-2) \cdots 4 \cdot 3$, and the like.

Returning to Lemma 2's proof, suppose a positive rational s/t is given. Use the sublemma to express $s(t-1)!$ in the form $d_1 t(t-1)(t-2) \cdots 4 \cdot 3 + d_2 t(t-1)(t-2) \cdots 5 \cdot 4 + \cdots + d_{t-3} t(t-1) + d_{t-2} t + d_{t-1}$. Dividing by $t!$ then yields a representation of s/t as a sum of reciprocals of factorials with appropriately small coefficients, which is the same as a terminating representation in the mixed-radix base \mathbf{b} .

Lemma 3. *The algorithm for digits of e is correct.*

Proof: It must be shown that $n+2$ mixed-radix digits of e suffice to get n base-10 digits of e . We first prove that if $n \geq 28$ ($= [10e]$), then n mixed-radix digits suffice for n base-10 digits. Using n mixed-radix digits means we are actually getting the base-10 digits of $e_n = (2; 1, 1, 1, \dots, 1) = \sum_{i=0}^n 1/i!$. Thus we must show that

$e - e_n \leq 5 \cdot 10^{-n}$ (see footnote at beginning of paper). A geometric series estimation of the tail of the series shows that $e - e_n < 2/(n+1)!$, and then Stirling's formula yields

$$\frac{2}{(n+1)!} < \frac{1}{n!} < \left(\frac{e}{n}\right)^n < \left(\frac{1}{10}\right)^n.$$

If $n < 28$ then a direct computation of the digits shows that $n+2$ mixed-radix digits suffice.

Lemma 4. *The base- c number with maximal digits, $(0; 2, 4, 6, 8, \dots)$, represents 2; hence regular representations of the form $(0; a, b, c, \dots)_c$ lie between 0 and 2.*

Proof: Instead of giving a formal proof, we show how some *Mathematica* computations led to the result (and a proof). In terms of series, the lemma states that

$$\sum_{i=0}^{\infty} \frac{(2i)i!}{(2i+1)!!} = 2.$$

A rough calculation showed that the sum is near 2. Then a rational computation of the remainders—the differences between the partial sums and 2—yielded the following sequence.

$$\frac{4}{3}, \frac{4}{5}, \frac{16}{35}, \frac{16}{63}, \frac{32}{231}, \frac{32}{429}, \frac{256}{6435}, \frac{256}{12155}, \frac{512}{46189}, \frac{512}{88179}.$$

The pattern in these remainders was found by dividing each by the preceding one, which yielded:

$$\frac{3}{5}, \frac{4}{7}, \frac{5}{9}, \frac{6}{11}, \frac{7}{13}, \frac{8}{15}, \frac{9}{17}, \frac{10}{19}, \frac{11}{21}.$$

Induction proves the pattern to be valid in general; it follows that the remainders have the closed form $2^{n+1}/\binom{2n+1}{n}$, which converges to 0, as claimed.

Lemma 5. *The algorithm for digits of π is correct.*

Proof: As for e , we look at $\pi - \pi_m$, where $\pi_m = (2; 2, 2, \dots, 2)_c$. This error is the tail of our main series for π : $\sum_{i=m}^{\infty} (i!)^2 2^{i+1} / (2i+1)!$. This tail is less than twice its first term since each subsequent term is less than half its predecessor, leading us to study $m!^2 2^{m+2} / (2m+1)!$. Splitting the denominator into evens and odds turns this into: $m! 2^2 / (3 \cdot 5 \cdots (2m+1))$, which is less than $\frac{2}{3} m! 2^2 / (2 \cdot 4 \cdots (2m))$, or $1/(3 \cdot 2^{m-1})$. It is easy to see (using the fact that $\frac{3}{10} < \log_{10} 2$) that this last is less than $5 \cdot 10^{-n}$ when $m = \lfloor 10n/3 \rfloor$, as claimed.

APPENDIX 2. PASCAL CODE

The following program, for which we are grateful to Macalester student Simeon Simeonov, implements the algorithm π -spigot. This code makes use of the fact that the queue of predigits always has a pile of 9s to the right of its leftmost member, and so only this leftmost predigit and the number of 9s need be remembered. The program computes 1000 digits of π and requires a version of Pascal with a longint data type (32-bit integer).

```

Program Pi_Spigot;
const n      = 1000;
len          = 10*n div 3;
var  i, j, k, q, x, nines, predigit : integer;
    a : array[1..len] of longint;
begin
  for j := 1 to len do a[j] := 2;           {Start with 2s}
  nines := 0; predigit := 0                {First predigit is a 0}
  for j := 1 to n do
    begin q := 0;
      for i := len downto 1 do              {Work backwards}
        begin
          x := 10*a[i] + q*i;
          a[i] := x mod (2*i-1);
          q := x div (2*i-1);
        end;
      a[1] := q mod 10; q := q div 10;
      if q = 9 then nines := nines + 1
      else if q = 10 then

```

```

begin write(predigit+1);
  for k := 1 to nines do write(0);           {zeros}
  predigit := 0; nines := 0
end
else begin
  write(predigit); predigit := q;
  if nines <> 0 then
    begin
      for k := 1 to nines do write(9);
      nines := 0
    end
  end
end;
writeln(predigit);
end.

```

ADDED IN PROOF. The latest version of *Mathematica* (2.3) can sum many of the series that occur in this paper. It takes only a second or so to get $\pi/2$ as the sum of the crucial series at the beginning of section 2, to get $1/i!$ for the series in Lemma 1's proof, and to get 2 as the sum of the series in Lemma 4's proof.

REFERENCES

- [Abd] S. Kamal Abdali, Algorithm 393—Special series summation with arbitrary precision, *Comm. ACM* **13** (1970) 570.
- [BBB] J. M. Borwein, P. B. Borwein, and D. H. Bailey, Ramanujan, modular equations, and approximations to pi, or How to compute one billion digits of pi, this *Monthly* **96** (1989) 201–219
- [Gos] R. W. Gosper, Acceleration of series, Memo no. 304, M.I.T. Artificial Intelligence Laboratory, Cambridge, Mass., 1974.
- [Knu] D. E. Knuth, *The Art of Computer Programming*, volume 2, Reading, Mass., Addison-Wesley, 1981.
- [Li] J. C. R. Li, Problem E854, this *Monthly* **56** (1949) 633–635.
- [Rab] S. Rabinowitz, Abstract 863-11-482: A spigot algorithm for pi, *Abstracts Amer. Math. Society* **12** (1991) 30.
- [Sale] A. H. J. Sale, The calculation of e to many significant digits, *Comput. J.* **11** (1968) 229–230.

MathPro Press
P.O. Box 713
Westford, MA 01886
72717.3515@compuserve.com

Department of Mathematics
Macalester College
St. Paul, MN 55105
wagon@macalstr.edu

The Rise, Fall, and Possible Transfiguration of Triangle Geometry: A Mini-history

Philip J. Davis

*For Deborah Tepper Haimo
In Friendship*

“Es ist in der That bewundernswuerdig, dass eine so einfache Figur, wie das Dreieck, so unerschoepflich an Eigenschaften ist. Wie viele noch unbekannte Eigenschaften anderer Figuren mag es nicht geben.” A.L. Crelle (1780-1855), *Sammlung*, v. I, 1821, p. 176.

[It is indeed wonderful that so simple a figure as the triangle is so inexhaustible in its properties. How many as yet unknown properties of other figures may there not be?]

1. INTRODUCTION. In the great *Encyklopaedie der Mathematischen Wissenschaften*, put out under the general editorship of Felix Klein, there will be found a hundred page article, completed in the Fall of 1914, on contemporary triangle geometry. (G. Berkhan and W. Fr. Meyer: *Neuere Dreiecksgeometrie*, Vol. III AB 10, pp. 1173-1276.) On finding this article, the reader’s eyebrows may be elevated: what on earth is triangle geometry? If the reader then goes to the index of the *Mathematical Reviews* for enlightenment, he will not find the term triangle geometry among the hundred or so subsets into which its coverage has been partitioned. Differential geometry, yes; convex geometry, yes; finite geometry, yes; triangle geometry, no. Yet, the *Encyklopaedie* devoted one of its major articles to this topic. F. Cajori, in his 1907 history of mathematics, devoted a half dozen pages to it. So what is going on here? The subsumption of this topic by another one? Or the essential death of a topic?

What, in fact, is triangle geometry? According to *Encyklopaedie* authors, it is not easy to define the subject logically, but it seems to boil down to this: given an arbitrary triangle, certain points (and lines and curves) are then determined which have remarkable properties with respect to the triangle. Instances of such points are the incenter, the circumcenter, the orthocenter, and the center of gravity of the triangle. By way of reminder, the first three are, respectively, the intersection of the internal angle bisectors, the intersection of the perpendicular bisectors of the triangle sides, and the intersection of the three altitudes. These four points were studied in antiquity.

In 1803, a mathematician by the name of Kluegel dubbed these points *the four distinguished (or remarkable) points* of the triangle (*merkwuerdige Punkte*). In the years that followed, a great many distinguished points, lines, circles, and conics, have been unearthed; so many, in fact, that Berkhan and Meyer despaired of counting them all. A point, line, circle, or conic, if sufficiently distinguished, merited a special name, and so we have, as some further examples, the Fermat

point, the Torricelli point, the Gergonne point, the Brocard points and circle, the Lemoine point and circle, the nine point circle, the Euler line, the symmedian point, the Steiner point, etc. etc. In Kimberling ("Central points...") will be found a listing of more than one hundred such distinguished objects. Not only special names were given, but also, quite understandably, histories of individual distinguished objects were written (e.g., Mackay wrote short histories of the symmedian point and of the nine-point circle. A recent book by Baptist presents many facts about the development of triangle geometry in the 19th century).

Accordingly, Berkhan and Meyer proposed as a definition of triangle geometry, "the study of distinguished points, lines, circles and conics of a triangle", leaving, as far as I can see, the definition of what is distinguished or remarkable about a point to one's subjective judgment.

A somewhat more sophisticated definition comes from Felix Klein himself (in his famous *Erlanger Programm*) and says that triangle geometry is the invariant theory of five points under the projective group. Perhaps this definition is less vague, but I don't think that it catches the flavor of the subject as it has been pursued historically.

Triangle geometry as a distinguished subfield of mathematics seems to have emerged in the 1870's in the writings of E. Lemoine, and if one considers the field both forward and backward in time from that date, it will be found that many distinguished mathematicians have contributed a little something to it. I leave to the reader's subjective judgment what constitutes a distinguished mathematician. Among the books that are wholly or partly devoted to this topic, one may cite Alasia (written under the encouragement of the famous geometer Eugenio Beltrami, then the President of the Reale Accademia dei Lincei), and containing, among other things, 566 metric formulas relating to the triangle and its distinguished points! Some other books are: Altschiller-Court, Casey, Coolidge, Emmerich, Johnson. In moments of euphoria, some of these authors viewed triangle geometry as the new and fulfilled Euclid, very much as the New Testament has been claimed as the fulfillment of the Old Testament.

One of the classic results of triangle geometry is the nine point circle theorem, which goes back, in part, to Poncelet in 1820. This theorem asserts that, given a triangle, the following nine points are concyclic: (i.e., lie on one circle) the three side bisectors, the three altitude feet, and the three midpoints along the altitudes from the vertices to the orthocenter. This is only one of the remarkable properties that this circle has; for example, it is tangent to the inscribed circle and the three escribed circles of the triangle. When one comes across this theorem in geometry for the first time, there is a certain surprise associated with it. One gets the feeling: what wonderful coincidences! However, there is nothing that dissipates such a feeling more quickly than to see the geometric theorem reduced to an algebraic identity or to have it placed in a more general context.

There is a great deal more that can be said about the nine-point circle. Some authorities have asserted that there are no fewer than 43 distinguished points lying on the nine-point circle. The nine-point circle theorem has generated a small mathematical industry. (See Gallatly). This, in itself, should now occasion no surprise, considering that a few well-chosen axioms such as those of group theory can generate a major mathematical industry. Moreover, it will give the reader an idea of the high regard accorded the nine-point circle to learn that some years ago, the distinguished analyst Dame Mary Cartwright told me that when she went up to Cambridge as a student (c. 1920), she was expected to know two different proofs of it.

One treatise on triangle geometry (Emmerich) presents the subject from the point of view of the Brocard points. I will refrain from giving definitions because the whole Brocard theory never sent me into raptures. But I shall mention one theorem that did; and when I was in high school, I cut my mathematical teeth on it.

“Napoleon’s Theorem”: On the sides of an arbitrary triangle T , erect three equilateral triangles outwardly. Then:

(1) The three centers of the equilateral triangle are themselves the vertices of an equilateral triangle. (Napoleon’s Triangle.)

(2) The three lines joining the vertices of the equilateral triangles to the opposite vertices of T are collinear. They meet in a point P known as the inner isogonic point of T . That is, P is the unique point in T at which the sides of T subtend equal angles of $2\pi/3$.

(3) These three line segments are of equal length.

(4) Similar statements when the equilateral triangles are constructed inwardly on the initial triangle.

(5) The inner and outer Napoleon Triangles have the same center.

(6) The areas of the inner and outer Napoleon triangles differ by the area of the initial triangle.

These are merely a few of the remarkable things that are associated with the Napoleon configuration or its generalizations. (See, e.g., Court, pp. 105-107, Sommerville, p. 165, Forder, p. 40, Hofstadter. And see Wetzel for a recent article containing new results and an extensive bibliography.)

I think that these examples should give the reader a good feeling for what triangle geometry is all about. I refer to the *Encyklopaedie* article of Berkhan and Meyer for a number of more complicated developments and to older bibliographical references. For additional recent references, see Kimberling.

How were the theorems of triangle geometry discovered? The mathematical literature, in general, is not often forthright as to how its material emerges. I can only conjecture that as with much of mathematics, it emerged from long hours of “playing around”. Playing around synthetically, in coordinate free fashion, the way that Euclid is written up; but also playing around with algebra, trigonometry, and rectangular, oblique, homogeneous, barycentric, trilinear, complex, conjugate, projective coordinates; all have been employed at one time or another. In textbooks such as Altshiller-Court, which are positioned as “advanced Euclid”, the synthetic approach is strong.

But another sort of playing around undoubtedly took place. The figures of triangle geometry can be drawn relatively easily and fairly accurately with a ruler and compass. I conjecture that a number of theorems were discovered visually in this way. Accurate computer graphics are now available for this sort of playing around, or to give this old and important activity its current gentrified name: mathematical experimentation.

2. TRIANGLE GEOMETRY BECOMES A MUSEUM PIECE. In a certain sense, the high regard accorded to triangle geometry culminated with the *Encyklopaedie* article. One of the two authors of the article (Berkhan) fell on the battlefield of World War I at the age of 32, his mathematical potential unrealized. As though prophetically, the subject itself hardly survived that war.

In the USA, triangle geometry was known as advanced geometry or college geometry. Courses were offered wherever there was a faculty devotee. Textbooks were written (e.g., N. Altshiller-Court, R. A. Johnson). In Sommerville’s 1924 book

on conics, a UK/New Zealand text, many theorems of triangle geometry were “downgraded” to the position of exercises for the student. In more recent years, a number of the theorems of triangle geometry appear in Coxeter, but there is no attempt to categorize them as such. The role of triangle geometry in European mathematical education has been detailed by Baptist.

Here is the 1940 judgment of Eric Temple Bell on the subject:

“The geometers of the 20th Century have long since piously removed all these treasures to the museum of geometry where the dust of history quickly dimmed their luster.” (*The Development of Mathematics*, p. 323)

Joseph Malkevitch, in a recent article that attempts to revive all kinds of geometry in the curriculum, lists fifty-eight subfields of geometry. Subfield No. 23, called Geometric Extremal Problems, lists the Fermat-Steiner Point. Other than that, there is nothing on triangle geometry on his list. Geometric interest, even when visual, turned elsewhere.

Yet, the subject of triangle geometry and its generalization, polygon geometry, (I often use the phrase triangle geometry to include this generalization), was and is still a steady source of problems for the entertainment and enjoyment of problem buffs who read the American Mathematical Monthly, the Mathematical Gazette, Crux Mathematicorum and similar periodicals in this and other countries. Over the years, one man, V. Thebault, contributed a thousand problems in the area.

Dozens of papers on these subjects have appeared—not just in the problem solving context—often displaying ingenious new approaches and new connections. Thus, Jesse Douglas presented a complex variable approach. I. J. Schoenberg exploited both complex variables and the discrete Fourier transform. Chang and Davis looked at Napoleon’s Theorem from the point of view of circulant matrices and the Moore-Penrose generalized matrix inverse. (Davis, 1977, 1979, Chang, Chang & Davis). Kimberling has examined triangle geometry from the point of view of functional equations; Baptist from the point of view of extremal problems.

Other than the quiet and steady problem solving activity and the occasional new result, the subject was making no waves. Mathematicians by and large, might play with individual items as a relaxation; they might even derive intense satisfaction, but they probably would not have wanted their professional reputation to be judged by a contribution of this sort.

“The song is ended, but the melody lingers on.”

3. WHY DID TRIANGLE GEOMETRY DIE? What reasons can be given for the short life of triangle geometry as a strongly and coherently delineated corpus of results, sanctioned by the mathematical establishment? I can suggest a few. Though hardly as complex a phenomenon as the decline and fall of the Roman Empire, I will not assert that I have gotten to the heart of the matter.

(1) The perception that the subject is part of elementary, “amateur”, or recreational mathematics and therefore is of low professional status. The subject is not “deep”. At the level of personal psychology, there was a feeling that even if a proof of a statement was not fairly transparent or immediately forthcoming, one could always “bulldoze” one’s way through a proof via analytic geometry. So why bother?

To speak of the professional status of certain problems, one must deal with the relationship between the inner challenges of a field and the outer sociology of

mathematicians. The latter includes the reward structure of mathematical activity. Where one group of mathematicians may charge a second group with “amateurism”, the second may counter with a charge of “elitism”. (See Fang and Takayama, Wilder, 1968, 1981). There is a similarity here with the field of music where the profession unselfconsciously divides its output into “classics”, “light classics”, “popular”, “highbrow”, “lowbrow”, and into many other status categories.

(2) The inner exhaustion of the interest and variety of its theorematic and methodologic possibilities. The *Encyklopaedie* article by Berkhan and Meyer presented no challenges or suggested directions for future development. No outstanding, long unsolved problems emerged to capture the imagination and challenge mathematical brilliance in the way that the famous Hilbert list of problems did. In a word, no really new ideas emerged from triangle geometry.

In this connection, however, I should mention one idea that emerged briefly from triangle geometry and is analogous to a currently thriving field: computational complexity. Called *geometrography*, it seems to have originated in a talk given by E. Lemoine at the Congress of the French Association for the Advancement of Science at Orano in 1888. The reader will find writeups in Alasia, pp. 29-44, in Coolidge, Chap. III, in Lemoine, and in Mackay (1893/4).

The idea of geometrography is as follows: beginning with a basic figure (often a triangle), construct a distinguished point or figure, often with ruler and compass, but also with other means, and then count the number of elementary operations required to do so.

Alasia’s elementary constructions (operations) are five in number (1) R_1 : Place a ruler’s edge through a given point. (2) R_2 draw a straight line. (3) C_1 : Place one point of a compass on a given point. (4) C_2 : Place one point of a compass on an indeterminate point of a line. (5). C_3 : Draw a circle. Now count up how many of these operations are required to effectuate the required construction. Call the total number of operations the *simplicity* of the construction.

The simplicity of many constructions will be found computed in the books alluded to. Here is one result: given the side of a regular pentagon, construct the circle in which the pentagon can be inscribed. The count given is $8R_1 + 4R_2 + 11C_1 + 8C_3$, yielding a coefficient of simplicity of 31.

Note again that the basic elementary operations are geometrical and not arithmetic. Even so, the coefficient of simplicity is strongly reminiscent of counting up the total number of floating point operations as is done in computer complexity theory.

As far as I am able to determine, the notion of constructive simplicity went nowhere in its day. It died on the vine.

(3) The increasing visual complexity and tediousness of the “deeper” results of triangle geometry, combined with

(4) A view of geometry that had emerged by the end of the 19th Century and seriously downgraded the visual in favor of the algebraic/symbolic.

(5) Susceptibility to the feeling of surprise has its ups and downs. (My goodness, do those three lines really intersect in one point? Who would have expected it?) But the professional is exposed to too many theorems and too many surprises. Surprise is accordingly dulled or attenuated and therefore devalued psychologically and can easily slip into boredom.

(6) The reassignment or the migration of some of the content of triangle geometry to other traditional or newly emerging fields. As examples: the famous Desargues theorem about two triangles in perspective is seen now as part of

projective geometry. Other theorems are viewed as part of inversive geometry or algebraic geometry.

(7) A dearth of connections or applications to other fields considered “live”; in particular, to areas in physics, etc.

However, some counterexamples exist here. Bernhard Neumann, who has written on triangle geometry (B. H. Neumann, 1941), told me that his father, Richard Neumann, who was an electrical engineer, discovered Napoleon’s Theorem on his own and made use of it in the theory of three phase alternating current circuits. (R. Neumann, 1911, 1939. B. H. Neumann, 1982.)

(8) Competition arising after World War II from many other geometrical constituencies, often with a strong visual component or with claims to wider applicability: e.g., convex geometry, tilings, symmetry and group theory, fractals, graph theory, computational geometry, etc. (See Malkevitch)

In a word, the problem of status boils down to this: for the reasons just outlined, and perhaps others, none of the major mathematicians of the post World War I period considered triangle geometry to be of great importance. It would be interesting to position its change in status within the context of the “laws” of the evolution of mathematics proposed by Crowe and by Wilder. (Wilder, 1968, 1981.)

4. ENTER: THE COMPUTER. It was clear, early on, that the computer offered the possibility of mathematical experimentation along visual, numerical, and symbolic/algebraic and logical lines; it offered the possibility of “mechanical” or “automatic” proof, and the possibility of the discovery or generation of new theorems.

With the availability of fast computation and convenient, high level languages, it was inevitable that the strategies directed toward the above goals would be applied to one of the “easiest” of the computable and decidable mathematical theories: good old triangle geometry.

The field of automated reasoning is currently extremely active, boasting of an enthusiastic corps of researchers, international conferences, and a number of specialized journals. Toward this end

*We may pursue a numerical road.

Let us suppose that a certain specific geometrical configuration has been put forward. Certain points, lines, curves, have been specified by their specific coordinates and certain conclusions are to be reached. Suppose that the conclusion may be reached by generation of a finite sequence of intersections of the curves coupled with the interpolation of new curves to currently available data. Carrying out such a program numerically, we reach our conclusion or verify our theorem in the specific numerical case set up. We may be even in a position to do all this visually using computer graphics.

In most instances, the numerical answers will be approximate; in favorable instances there will be single or multiple precision accuracy. By altering the numerical parameters, a family of results can be displayed rapidly, and on this basis certain distinguished phenomena, occurrences, (theorems) may be inferred by the investigator. This kind of thing goes on constantly in computer graphics, or in computer assisted geometric industrial design (CAGD).

The investigator whose criterion of validity demands more than approximate numerical verification in one specific instance must employ other strategies. If the given initial configuration consists solely of lines specified as passing through points with rational coordinates, then all the computation may be (in principle) performed exactly in rational arithmetic.

There is a strategy available to overcome numerically the lack of generality in one special numerical case. It will be exhibited in the case of the famous Pappus theorem of projective geometry whose initial configuration is two arbitrary straight lines and three arbitrary points on each.

If the coordinates of the points and lines are taken as algebraically independent real numbers, then the verification of the theorem *in that one case* serves to demonstrate the theorem generally. (Davis, 1977, Rowland and Davis, 1981, 1981, Schwartz, Hong and Tan).

This leaves hanging in the air the question of what, digitally speaking, a set of algebraically independent numbers could possibly mean. Such a set would act as our symbols with respect to the computations involved, and hence the force of the above remark.

One might instantiate a “pseudo-algebraically independent” set of numbers by taking them as random numbers, and interpreting the result probabilistically. One then arrives at the following principle: (valid within a certain limited context) if a theorem is true for one randomly selected set of initial configurations, it is true for all configurations. This puts at risk the old caveat of mathematics teachers (which can have a constipating effect on investigations): you must prove it in *all* cases, not just in one special case. At the same time, it offers the possibility in some instances of formalizing the inductive leaps that the mathematical mind takes when confronted with what seems to be, logically speaking, incomplete evidence.

*We may go the road of heuristics, such as developed by George Pólya in a series of popular books. We may try to combine Pólya heuristics with strategies of Artificial Intelligence (AI). (Newell. R. Davis and D. Lenat.)

*We may go the road of logic. Tarski has proved that all statements in “Tarski geometry” are decidable. But it has been found that working with the two basic predicates given by Tarski, one for betweenness and one for distance, has not been a very promising approach. (Woos, p. 206-214, Hao Wang. For a system of “natural deduction” using the program AUTOMATH, see de Bruijn).

*We may travel along the symbolic road, using computer packages such as FORMAC (now located in the Museum of Ancient Software!), MAPLE, MACSYMA, MATHEMATICA, and prove Pappus, in a naive and ad hoc way. (Davis and Cerutti)

*We may travel a rather sophisticated algebraic road, a road traveled by Wu, Chou, Zhang, Goa, and others, a road that uses such algebraic ideas as Ritt’s Principle, or Groebner Bases.

Briefly, and here I follow Chou, 1988, the strategy of Wu’s method is:

Step 1: First convert the initial geometrical configuration into a set of polynomial equations. Convert the geometrical conclusion into a polynomial equation.

The initial configuration (hypotheses) will be specified by

$$h_1(u_1, u_2, \dots, u_d; x_1, \dots, x_t) = 0.$$

$$h_2(u_1, u_2, \dots, u_d; x_1, \dots, x_t) = 0.$$

$$h_n(u_1, u_2, \dots, u_d; x_1, \dots, x_t) = 0.$$

The conclusion is given by

$$g(u_1, u_2, \dots, u_d; x_1, \dots, x_t) = 0.$$

In these equations, the u 's are independent variables, while the x 's are algebraically dependent on the u 's.

Step 2. Using pseudo division and Ritt's Principle, (or your own method), "triangulate" the polynomials. This means replacing the set of h 's by a set which introduces one new x at a time. Then check for irreducibility.

Step 3. Successive pseudodivision to arrive at a final remainder R , after analysis of certain non-degenerate conditions. Hopefully, $R = 0$, indicating that the geometrical implication is true.

To indicate the complexity of this approach in some specific examples, Chou reports that the proof of the so-called Thebault-Taylor theorem of elementary geometry (involving lines, circles, intersections and tangencies) required the manipulation of polynomials of almost 700,000 terms. (In their naive approach to Pappus' Theorem of projective geometry, Davis and Cerutti report that the number of terms in the polynomials was almost 33,000.)

5. GENERATION OF NEW THEOREMS USING THE COMPUTER. A number of approaches have been explored.

*Playing around (i.e., mathematical experimentation) visually, numerically, symbolically. This is quite successful. For example, I have found many theorems (unpublished) in the area of group matrices simply by playing around with the MATLAB matrix package. Most experimenters can report similar experiences.

In connection with visual output, I have even argued (Davis, 1974), for the recognition of the existence of "visual theorems", i.e., stable visual patterns, generated by a computer algorithm, where what the eye "sees" need not even be verbalized, let alone formalized in traditional mathematical language. (As a parallel, philosopher Susanne Langer, in the context of music, speaks of the "subtle complexes of feeling that language cannot even name, let alone set forth".)

*Programmed heuristics. This seems to be less promising. (See, e.e., Newell, 1981) R. Davis and D. Lenat have written a program, AM (Automated Mathematician) which starts from set theory, and proposes to invent new mathematical concepts and new conjectures relying on a library of built-in heuristics.

Whatever the method employed, numerous new theorems have emerged, some of which have attracted attention, surprise and enthusiasm. (Hofstadter, Grünbaum and Shephard. In the last named reference, there are some philosophical remarks on proof methodology paralleling those just made.)

6. THE TRANSFIGURATION OF TRIANGLE GEOMETRY. Can a subject arise from the dust and ashes that history has piled on it? Only if it is transformed in the process. The focus of triangle geometry has now been changed. The computer has popped it up a metalevel, and in the process has transfigured the subject. Hundreds of elementary and not so elementary theorems that were in the literature have now been proved by computer. Many new theorems have been discovered, again in a variety of ways. Triangle geometry always was a practice ground for strategies of proof in the spirit of Euclid, and it has now become a testing ground for strategies of decidability, proof, and theorem discovery. These strategies have run from naive schemes to the employment of deep and abstract results of modern algebra and differential algebra.

But there is yet more that emerges from the change of focus: I believe that the experience gained in this change can become a prime source of raw material for

philosophical discussions on the nature of proof, methodologies of research, the role and nature of intuition, educational values, etc.

What are some of the implications of this work? While I think it is too early to write with assurance, I will venture a few observations.

*Obsessed over the millennia by the vision that mathematics can provide absolute, rock bottom “certainty”, the mathematical establishment has often expressed its displeasure with certain types of “proof”: visual, mechanical, experimental, probabilistic. This attitude goes back as far as Archimedes (200 B.C.), if not further.

Computer proof, theorem discovery, and mathematical experimentation are now openly acknowledged as legitimate methodologies and roads to mathematical knowledge.

Thus, absolutely rigorous mathematical proof, as an ideal, is giving way and is now seen as a part of a wider, more generous and more flexible notion that I like to call “mathematical evidence”.

*Given that the output of the whole mathematical world, measured in terms of numbers of theorems, is of the order of one hundred thousand per annum, what use would the automatic generation of theorems in a restricted, well ploughed area serve? As A. L. Crelle correctly observed in the epigraph placed at the head of this article, the simplest mathematical structure can produce an unlimited number of conclusions. What, therefore, does one “do” with mathematical products that might be stamped out like doughnuts in a doughnut machine?

*In this process, the individual theorem may stand devalued. For example, it may be vital in a certain application to know that the product $12563 \times 502 = 6306626$, but to the average mathematical mind, the theorem expressed by this multiplicative identity is quite tedious.

Since this kind of theorem (of arithmetic, triangle geometry or whatever) can now be produced by the hundredfold, the emphasis inevitably changes from the theorem to the means by which the theorem is produced. By and large, the medium becomes the message. This is one of the lessons taught by the subsumption of geometry by algebra that occurred as a result of the revolutionary vision of Descartes.

*An individual theorem may still be judged as to its importance, practical or otherwise, and may turn out to have such importance. This importance is determined by subjective and historical criteria.

The process whereby a mathematical concept, whether it be a simple point in a triangle or a whole complex theory, becomes “distinguished”, is not capable of formalization. (Woos sets up this problem as one of 33 basic research problems in automated reasoning.) It is a historical process and may involve the whole scientific community or significant subsets of that community.

*The inner complexity of some proofs by computer, often involving polynomials of hundreds or thousands of terms, adds new respect and appreciation for the historic methods and traditional manner in which the results were presented.

*The mysterious, omnipresent and vitally essential “mathematical intuition” together with its components of experience, analogy, educated guessing, and transcendental, non-explainable “pre-knowledge” (e.g., Ramanujan), all get raised a metalevel and now can operate in a wider arena.

*As regards mathematical education, I think the message is clear. Classical proof must move over and share the educational stage and time with other means of arriving at mathematical evidence and knowledge. Mathematical textbooks must modify the often deadening rigidity of the Euclidean model of exposition.

7. CONCLUDING REMARKS.

“The whole cultural world, in all its forms exists through tradition. These forms have arisen not merely causally . . . they have arisen within our human space through human activity.” —Edmund Husserl, *The Origin of Geometry*.

The distinguished logician Hao Wang, as a result of his interest in proof by computer, was once charged with desiring to eliminate the mathematician. He answered “No, only the inferior ones”. I should like to interpret this remark more charitably: that the capabilities of all mathematicians are elevated by their association with computation. The transformation by the computer of triangle geometry, and of many other areas has, paradoxically, reconfirmed and strengthened the vital role of humans in the wonderful activity known as doing mathematics. Put it even more strongly: mathematics develops in such a way that the role of the mathematicians is always manifest.

ACKNOWLEDGMENTS. The author wishes to acknowledge helpful suggestions received from Drs. Christa Binder, Branko Grünbaum, Douglas Hofstadter and from the referees.

BIBLIOGRAPHY

- Alasia, Christoforo, *La Recente Geometria del Triangolo*, Citta di Castello: S. Lapi, 1900.
- Altshiller-Court, N. *College Geometry*, Richmond: Johnson, 1925.
- Baptist, Peter, *Die Entwicklung der Neueren Dreiecksgeometrie*, Mannheim: B. I. Wissenschaftsverlag, 1992.
- Bell, Eric T., *The Development of Mathematics*, New York: McGraw Hill, 1940.
- Berkhan, G. and W. Fr. Meyer: Neuere Dreiecksgeometrie, *Encyklopaedia der Mathematischen Wissenschaften*, Leipzig: Teubner, 1914–1931, Vol. III, AB 10, pp. 1173–1276.
- Bledsoe, W. W. and D. W. Loveland, eds. *Automated Theorem Proving: After 25 years*, Providence: American Math. Soc., 1984.
- de Bruijn, N. G., A survey of the project Automath. In: To H. B. Curry: *Essays in combinatory logic, lambda calculus and formalism*, ed. J. P. Seldin and J. R. Hindley, New York: Academic Press 1980.
- Cajori, F. *A History of Elementary Mathematics*, New York: Macmillan, 1917.
- Casey, John, *Geometrie Elementaire Récente*, Paris: Gauthier-Villars, 1890.
- Chang, Geng-Zhe, A proof of the theorem of Douglas and Neumann by circulant matrices, *Houston Math. J.*, vol 8, 1982, pp. 15–18.
- Chang, Geng-Zhe, and P. J. Davis, A circulant formulation of the Napoleon-Douglas-Neumann Theorem, *Linear Algebra and its Applications*, 1983, pp. 87–95.
- Chou Shang-Ching, An Introduction to Wu’s methods for mechanical theorem proving in geometry, *Journal of Automated Reasoning*, vol 4, 1988, pp. 237–267.
- Coolidge, Julian L., *A Treatise on the Circle and the Sphere*, Oxford: Clarendon Press, 1917. Reprinted, Bronx, New York: Chelsea, 1971.
- Coxeter, H. S. M., *Introduction to Geometry*, New York: Wiley, 1961.
- Coxeter, H. S. M. and S. L. Greitzer, *Geometry Revisited*, Washington: Math. Assn. Amer, 1967.
- Crowe, M. J., Ten ‘laws’ concerning patterns of change within the history of mathematics, *Historia Mathematica*, v. 2, pp. 161–166. Also pp. 469–470.
- Davis, P. J., Visual Geometry, Computer Graphics and Theorems of Perceived Type, *Proc. Symp. Appl. Math.* vol. 20, Amer. Math. Soc., 1974.
- , Cyclic transformations of polygons, and the generalized inverse, *Can. J. Math.*, 29, pp. 756–770, 1977.
- , Proof, completeness, transcendentals and sampling, *J. Assoc. Comp. Mach.*, vol. 24, 1977, pp. 298–310.
- , *Circulant Matrices*, New York: Wiley, 1979, pp. 140–154.
- , Are there coincidences in mathematics?, *Amer. Math. Month.*, vol. 88, 1981, pp. 311–320.
- Davis, P. J., and Elsie Cerutti, FORMAC meets Pappus: Some observations on elementary analytic geometry by computer, *Amer. Math. Monthly*, vol. 76, 1969, pp. 895–905.
- Davis, Randall and Douglas Lenat, eds. *Knowledge Based Systems in Artificial Intelligence*, New York: McGraw-Hill, 1982.

- Douglas, Jesse, Geometry of Polygons in the Complex Plane, J. Math. and Phys., vol. 19, 1940, pp. 93–130.
- Emmerich, A., *Die Brocardschen Gebilde und ihre Beziehungen zu den Verwandten Merkwuerdigen Punkten und Kreisen des Dreiecks*, Berlin: Georg Reimer, 1891.
- Fang, J. and K. P. Takayama, *Sociology of Mathematics and Mathematicians*, Hauppauge, N.Y.: Paideia, 1973.
- Forder, H.G., *The foundations of Euclidean Geometry*, Cambridge: University Press, 1927.
- Gale, David, Mathematical Entertainments, The Mathematical Intelligencer, Spring, 1992.
- Gallatly, William, *The Modern Geometry of the Triangle*, London: 1910.
- _____, *The Nine Point Circle*, Cambridge: Johnson, 1907.
- Grünbaum, Branko and G. C. Shephard, From Menelaus to Computer Assisted Proofs in Geometry, to appear.
- Hofstadter, Douglas R., From Euler to Ulam: discovery and dissection of a geometric gem, Center for Research on Concepts and Cognition, Bloomington, Indiana, December, 1992.
- Hong, Jiawei and Xiao-nan Tan, Proving inequalities by examples, Courant Institute of Math., 1987.
- Johnson, R. A. *Advanced Euclidean geometry*, New York: Dover, 1960.
- Kimberling, Clark H. Central points and central lines in the plane of a triangle. To appear in Mathematics Magazine.
- Kimberling, Clark C., Triangle centers as functions. To appear.
- _____, Functional equations associated with triangle geometry. Aequationes Mathematicae.
- _____, and Peter Yff, The circumcircle and the line at infinity. To appear.
- Klee, Victor and Stan Wagon, *Old and new unsolved problems in plane geometry and number theory*, Washington: Math. Assos. Amer., 1991.
- Lemoine, E., *Géométophraphy*, Paris: 1902.
- Mackay, J. S., History of the Nine Point Circle, Proc. Edin. Math. Soc., v. 11, 1892-3, pp. 19–57.
- _____, Early history of the symmedian point, Proc. Edin. Math. Soc., v. 11, 1892-3, pp. 92–103.
- _____, The Geometography of Euclid's Problems, Proc. Edin. Math. Soc., v. 12, 1893-4, pp. 2–16.
- Malkevitch, Joseph, Geometry in Utopia, Dept of Math., York College, Jamaica, N.Y. 11451, March 3, 1989.
- Neumann, Bernhard H., Some remarks on polygons, J. Lond. Math. Soc., vol. 16, 1941, pp. 230–245.
- _____, Plane polygons revisited, pp. 113–122 of *Essays in Statistical Science: Papers in Honour of P.A.P. Moran*, J. Gani and E. J. Hannan, eds., J. of Applied Probability, Special Volume 19A, 1982.
- Neumann, Richard, Geometrische Untersuchung eines Ausgleichstransformators fuer unsymmetrische Drehstromsysteme, Elektrotechnik und Maschinenbau, vol. 39 1911, pp. 747–751.
- _____, *Symmetrical Component Analysis*, London: Pitman: 1939.
- Newell, Allen, The Hueristic of George Pólya and its relation to artificial intelligence, Dept. Comp. Sci., Pub. No. 81–133, Carnegie-Mellon Univ. Pittsburgh, 1981.
- Pedoe, Daniel, *Circles*, London: Pergamon, 1957.
- Rowland, John H. and P. J. Davis, On the selection of test data for recursive mathematical subroutines, SIAM J. Comput. vol. 10, 1981, pp. 59–72.
- _____, and _____, On the use of transcendentals for program testing, J. Assoc Comp. Mach., vol. 28, 1981, pp. 181–190.
- Schoenberg, I. J., *Mathematical Time Exposures*, Washington: Math. Assn. of Amer., 1982. Chaps. 6–9.
- Schwartz, J. T., Probabilistic Algorithms for Verification of Polynomial Identities, Journal of the ACM, 1980.
- Sommerville, D. M. Y., *Analytical Conics*, London: G. Bell, 1924.
- Tarski, A., *A decision method for elementary algebra and geometry*, 2nd. ed. Berkeley: 1951.
- Wang, Hao, Computer theorem proving and artificial intelligence, pp. 49–72 of Bledsoe and Loveland.
- Wetzel, John E., Converses of Napoleon's Theorem, Amer. Math. Monthly, vol. 99, 1992, pp. 339–351.
- Wilder, Raymond L., *Mathematics as a Cultural System*, Oxford: Pergamon Press, 1981
- _____, *The Evolution of Mathematical Concepts*, New York: Wiley, 1968.
- Woos, Larry, *Automated Reasoning: 33 Basic Research Problems*, Englewood Cliffs: Prentice Hall, 1987.
- Wu Wen-Tsuen, Basic principles of mechanical theorem proving in geometries, Journal of Automated Reasoning, v. 2, 1986, pp. 221–252.

Division of Applied Mathematics
Brown University
Providence, RI 02912

Totally Real Origami and Impossible Paper Folding

David Auckly and John Cleveland

Origami is the ancient Japanese art of paper folding. It is possible to fold many intriguing geometrical shapes with paper [M]. In this article, the question we will answer is which shapes are possible to construct and which shapes are impossible to construct using origami. One of the most interesting things we discovered is that it is impossible to construct a cube with twice the volume of a given cube using origami, just as it is impossible to do using a compass and straight edge. As an unexpected surprise, our algebraic characterization of origami is related to David Hilbert's 17th problem. Hilbert's problem is to show that any rational function which is always non-negative is a sum of squares of rational functions [B]. This problem was solved by Artin in 1926 [Ar]. We would like to thank John Tate for noticing the relationship between our present work and Hilbert's 17th problem. This research is the result of a project in the Junior Fellows Program at The University of Texas. The Junior Fellows Program is a program in which a junior undergraduate strives to do original research under the guidance of a faculty mentor.

The referee mentioned two references which the reader may find interesting. "Geometric Exercises in Paper Folding" addresses practical problems of paper folding [R]. Among many other things, Sundara Row gives constructions for the 5-gon, the 17-gon, and duplicating a cube. His constructions, however, use more general folding techniques than the ones we consider here. Felix Klein cites Row's work in his lectures on selected questions in elementary geometry [K].

In order to understand the rules of origami construction, we will first consider a sheet of everyday notebook paper. Our work with notebook paper will serve as an intuitive model for our definition of origami constructions in the Euclidean plane. There are four natural methods of folding a piece of paper. The methods will serve as the basis of the definition of an origami pair.

We construct the line L_1 , by folding a crease between two different corners of the paper. Another line may be constructed by matching two corners. For example, if corners α and γ are matched, the crease formed, L_2 , will be the perpendicular bisector of the segment $\alpha\gamma$. Another natural construction is matching one line to another line. For instance, $\beta\gamma$, the paper's edge, and L_2 are lines. If we lay $\beta\gamma$ upon L_2 and form the crease, then we obtain L_3 which is the angle bisector of the two lines. If we start with two parallel lines in this third construction, then we will just get a parallel line half way in between.

The fourth and final construction which seems natural is consecutive folding. This is similar to rolling up the sheet of paper only one does not roll it up, he folds it up. More explicitly, start with a piece of paper with two creases on it as in Figure 2. Fold along line L_1 and do not unfold the piece of paper. Notice that line L_2 lies

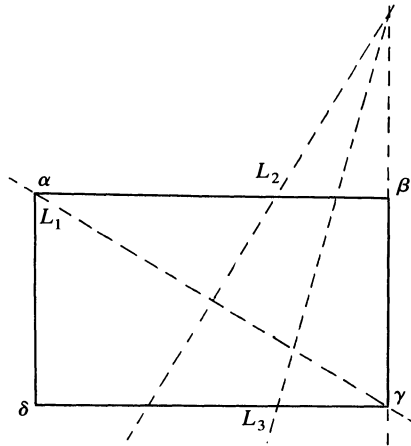


Figure 1

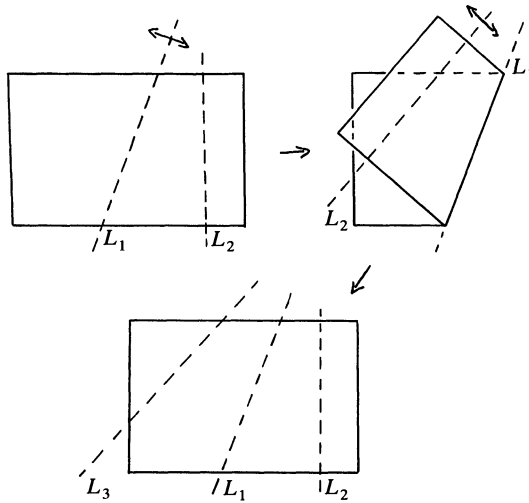


Figure 2

over the sheet of paper. With the paper still folded, fold the sheet of paper along the crease L_2 to obtain a new crease on the sheet underneath L_2 . If we name this new crease L_3 and unfold the sheet of paper, then it is easy to see that line L_3 is the mirror image or reflection of line L_2 about line L_1 .

We now formalize these methods to define an origami pair on the plane. The creases on our sheet of paper are merely lines in the plane, and the corners of the paper are represented by points where lines (creases) meet. This previous discussion is the motivation for the following definition.

Definition. $(\mathcal{P}, \mathcal{L})$ is an *origami pair* if \mathcal{P} is a set of points in \mathbb{R}^2 and \mathcal{L} is a collection of lines in \mathbb{R}^2 satisfying:

- i) The point of intersection of any two non-parallel lines in \mathcal{L} is a point in \mathcal{P} .
- ii) Given any two distinct points in \mathcal{P} , there is a line in \mathcal{L} going through them.

- iii) Given any two distinct points in \mathcal{P} , the perpendicular bisector of the line segment with given end points is a line in \mathcal{L} .
- iv) If L_1 and L_2 are lines in \mathcal{L} , then the line which is equidistant from L_1 and L_2 is in \mathcal{L} .
- v) If L_1 and L_2 are lines in \mathcal{L} , then there exists a line L_3 in \mathcal{L} such that L_3 is the mirror reflection of L_2 about L_1 .

For any subset of the plane containing at least two points, there is at most one collection of lines which will pair with it to become an origami pair.

Definition. A subset of \mathbb{R}^2 , \mathcal{P} , is *closed under origami constructions* if there exists a collection of lines, \mathcal{L} , such that $(\mathcal{P}, \mathcal{L})$ is an origami pair.

The question which we answer in this paper is which points may be constructed from just two points, using only the origami constructions described above. We will call that collection of points the set of origami constructible points.

Definition. $\mathcal{P}_0 = \cap \{\mathcal{P} | (0, 0), (0, 1) \in \mathcal{P} \text{ and } \mathcal{P} \text{ is closed under origami constructions}\}$ is the set of *origami constructible points*.

Before we explain the structure of \mathcal{P}_0 , we give an example of an origami construction analogous to many compass and straight edge constructions, namely, the construction of parallel lines.

Lemma. *It is possible to construct a line parallel to a given line through any given point using origami.*

Proof: Refer to Figure 3. Given a line L and a point p , pick two points p_1 and p_2 on L . By property ii) in the definition of an origami pair, we may construct lines L_1 and L_2 running through p_1, p and p_2, p , respectively. By property v) we may reflect L_1 and L_2 through L to obtain L_3 and L_4 . Now the intersection of L_3 and L_4 is a constructible point, so there is a line, L_5 through this point and the given point, p , by properties i) and ii). Call the point where L_5 and L intersect p_3 . To finish the construction, use property iii) to construct a perpendicular bisector to p, p_3 , and reflect L through this bisector with property v) to obtain the desired line, L_6 . It is a straightforward exercise to show that L_6 has the desired properties.

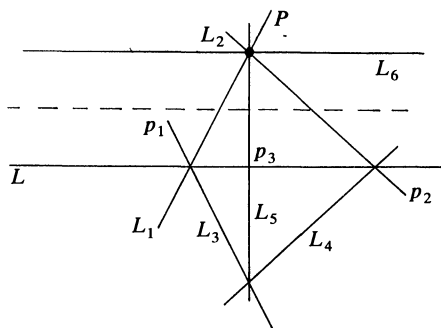


Figure 3

The reader may wish to try some constructions on his own. Two especially interesting exercises to attempt are the construction of a right triangle with given legs and the construction of a right triangle with a given hypotenuse and leg. More explicitly, given four distinct points α, β, γ and δ , the reader may try to construct a point ε such that $\alpha, \beta, \varepsilon$ are the vertices of a right triangle with legs $\overline{\alpha\beta}$ and $\overline{\beta\varepsilon}$ such that the length of $\overline{\beta\varepsilon}$ equals the length of $\overline{\gamma\delta}$.

Now that we have a better feel for origami constructions, we will start developing tools to show that some figures are not constructible. The first thing we need is the notion of an origami number.

Definition. $\mathbb{F}_0 = \{\alpha \in \mathbb{R} | \exists v_1, v_2 \in \mathcal{P} \text{ such that } |\alpha| = \text{dist}(v_1, v_2)\}$ is the *set of origami numbers*.

It is easy to see that $(x, y) \in \mathcal{P}_0$ if and only if x and y are both in \mathbb{F}_0 . It is also easy to see that the numbers $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \dots$ are origami numbers. To see that $\frac{1}{5}$ is an origami number consider Figure 4. In Figure 4 a line through $(0, \frac{5}{8})$ and $(1, 0)$ is constructed, then a parallel line through $(0, \frac{1}{8})$ is constructed. This parallel line intersects the x -axis at $(\frac{1}{5}, 0)$, therefore $\frac{1}{5}$ is an origami number. Another class of origami numbers can be generalized by a simple geometric construction. Starting with any segment, it is possible to construct a right triangle as in Figure 5. It follows that $\sqrt{1 + \alpha^2}$ is an origami number whenever α is an origami number. Using this construction, we see that

$$\sqrt{2} = \sqrt{1 + 1^2} \quad \text{and} \quad \sqrt{3} = \sqrt{1 + (\sqrt{2})^2}$$

are origami numbers. In fact, the sum, difference, product, and quotient of origami numbers are origami numbers.

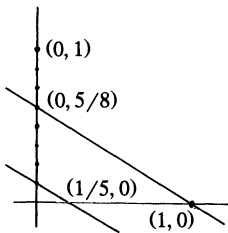


Figure 4

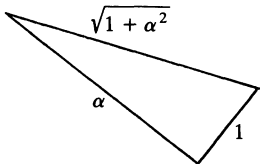


Figure 5

Theorem. The collection of origami numbers, \mathbb{F}_0 is a field closed under the operation $\alpha \mapsto \sqrt{1 + \alpha^2}$.

Proof: If $\alpha, \beta \in \mathbb{F}_0$, it follows from the definition that $-\alpha \in \mathbb{F}_0$ and it is easy to show that $\alpha + \beta \in \mathbb{F}_0$. Straightforward constructions with similar triangles are enough to show that $\alpha \cdot \beta, \alpha^{-1} \in \mathbb{F}_0$. See Figure 6. In the discussion preceding this theorem we showed that $\sqrt{1 + \alpha^2}$ is an origami number whenever α is. The proof is therefore complete.

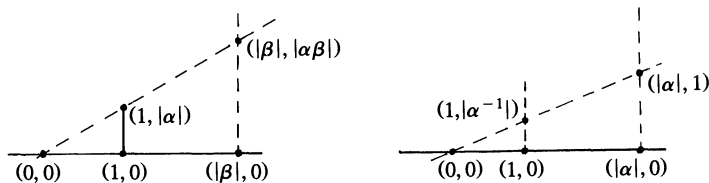


Figure 6

Now that we have some algebraic operations which will produce origami numbers, it is natural to ask if there are any more operations which will produce origami numbers. Once we have a list of all ways to create origami numbers and a method to test if a given number can be achieved, then we will know which geometric shapes are constructible, and which shapes are not constructible. This is because any figure is constructible if and only if the coordinates of all of the vertices are origami numbers.

Definition. $\mathbb{F}_{\sqrt{1+x^2}}$ is the smallest subfield of \mathbb{C} closed under the operation $x \mapsto \sqrt{1 + x^2}$.

The preceding Theorem may be rephrased as $\mathbb{F}_{\sqrt{1+x^2}} \subset \mathbb{F}_0$. It is in fact true that $\mathbb{F}_0 = \mathbb{F}_{\sqrt{1+x^2}}$. Thus, the previously listed operations which produce origami numbers are the only independent operations which produce origami numbers.

Theorem. $\mathbb{F}_0 = \mathbb{F}_{\sqrt{1+x^2}}$.

Proof: Since we already know that $\mathbb{F}_{\sqrt{1+x^2}} \subset \mathbb{F}_0$, we only need to show that $\mathbb{F}_0 \subset \mathbb{F}_{\sqrt{1+x^2}}$. That is, we need to show that any origami number may be expressed using the usual field operations and the operation $x \mapsto \sqrt{1 + x^2}$. It is enough to consider the coordinates of origami constructible points, because a number is an origami number if and only if it is a coordinate of a constructible point. There are only four distinct ways of constructing new origami points from old ones using the axioms for origami construction. These are illustrated in Figures 7 and 8. The only way a new point will be constructed is by a new crease intersecting an old one. The four ways of making a crease are: folding a line between two existing points as in the line $\gamma\delta$ in Figure 7, folding the perpendicular bisector to two points as in the second part of Figure 7, reflecting a line as in the third part of Figure 7, or forming

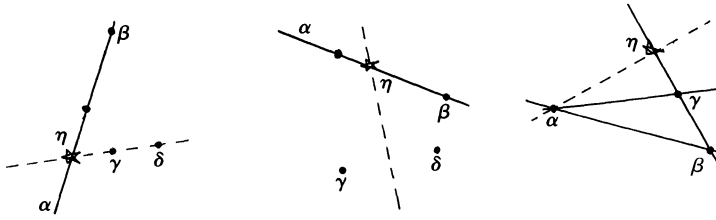


Figure 7

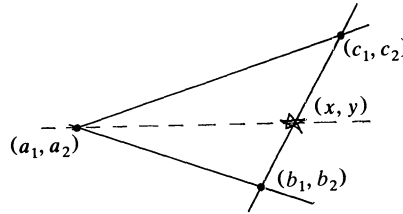


Figure 8

the angle bisector as in Figure 8. We will explain the case illustrated in Figure 8 and leave the remaining three cases to the reader. When showing that the point (x, y) only depends on the prescribed operations, we may assume that $(a_1, a_2) = (0, 0)$ by translation, because the point (x, y) is found by adding (a_1, a_2) to the translated point. We may further assume that (b_1, b_2) is on the unit circle, by scaling because multiplying by $\sqrt{b_1^2 + b_2^2} = |b_1|\sqrt{1 + (b_2/b_1)^2}$ will reverse the scaling. Even further (b_1, b_2) may be assumed to be $(1, 0)$ because the rotation $(x, y) \mapsto (b_1x - b_2y, b_2x + b_1y)$ sends the point $(1, 0)$ back to (b_1, b_2) . Let $\theta = \angle cab$, with the above assumptions, $\cot \theta = c_1/c_2$ and

$$\csc \theta = \sqrt{c_1^2 + c_2^2}/c_2 = \sqrt{1 + (c_1/c_2)^2}.$$

Now the slope of the new crease is $m = \tan(\theta/2) = \csc \theta - \cot \theta$, which only depends on the prescribed operations. The new point (x, y) is the intersection of the two lines $y = mx$ and $y = [c_2/(c_1 - 1)](x - 1)$, so

$$x = \frac{c_2}{c_2 - m(c_1 - 1)} \quad \text{and} \quad y = \frac{mc_2}{c_2 - m(c_1 - 1)}$$

which only depends on the prescribed operations as was to be shown.

The preceding theorem gives an algebraic description of the field of origami numbers, and in principle answers which shapes are constructible and which are not constructible using origami. In practice it is still difficult to decide whether or not a given number is an origami number. For example, $\sqrt{4 + 2\sqrt{2}}$ is an origami number because $\sqrt{4 + 2\sqrt{2}} = \sqrt{1 + (1 + \sqrt{2})^2}$, but what about $\sqrt{1 + \sqrt{2}}$? In order to answer this question we need a better characterization of origami numbers. Before we proceed we will review some elementary facts from abstract algebra [AH], [L].

Definition. A number, α , is an *algebraic number* if it is a root of a polynomial with rational coefficients.

Any algebraic number, α , is a root of a unique monic irreducible polynomial in $\mathbb{Q}[x]$, denoted by $p_\alpha(x)$. This polynomial, moreover, divides any polynomial in $\mathbb{Q}[x]$ having α as a root.

Definition. The *conjugates* of α are the roots of the polynomial $p_\alpha(x)$. An algebraic number is *totally real* if all of its conjugates are real. We denote the set of totally real numbers by \mathbb{F}_{TR} .

Of the numbers which we are using to motivate this section, $\sqrt{4 + 2\sqrt{2}}$ is totally real, because all of its conjugates ($\pm \sqrt{4 \pm 2\sqrt{2}}$) are real, but $\sqrt{1 + \sqrt{2}}$ is not totally real because two of its conjugates are imaginary ($\pm \sqrt{1 - \sqrt{2}}$).

The last topic which we review is symmetric polynomials. The symmetric group on n letters acts on polynomials in n variables by $\sigma f(x_1, x_2, \dots, x_n) = f(x_{\sigma(1)}, x_{\sigma(2)}, \dots, x_{\sigma(n)})$ where $f \in R[x_1, \dots, x_n]$ and R is an arbitrary ring.

Definition. The fixed points of the above action are called *symmetric polynomials* over R .

For example, $x_1^2 + x_2^2$ is a symmetric polynomial in two variables because it remains unchanged when the variables are interchanged. However, $x_1^2 - x_2^2$ is not a symmetric polynomial because it becomes $x_2^2 - x_1^2 \neq x_1^2 - x_2^2$ when x_1 and x_2 are interchanged. One important class of symmetric polynomials is the class of elementary symmetric polynomials.

Definition. If $\prod_{k=1}^n (t + x_k)$ is expanded, we obtain

$$\prod_{k=1}^n (t + x_k) = \sum_{l=0}^n \sigma_l(x_1, \dots, x_n) t^{n-l}.$$

The $\sigma_l(x_1, \dots, x_n)$ are the *elementary symmetric polynomials*.

It is easily verified that

$$\sigma_1 = x_1 + x_2 + \dots + x_n$$

$$\sigma_l = \text{the sum of all products of } l \text{ distinct } x_k \text{'s}$$

$$\sigma_n = x_1 \cdot x_2 \cdot \dots \cdot x_n.$$

Fact. [L, page 191]. The algebra of symmetric polynomials over R is generated by the elementary symmetric polynomials. That is, any symmetric polynomial is a linear combination of products of the elementary symmetric polynomials.

We will now begin the final characterization of the origami numbers. It happens that all origami numbers are totally real. To prove this, it is necessary to show that the sum, difference, product and quotient of totally real numbers is totally real, and that $\sqrt{1 + \alpha^2}$ is totally real whenever α is totally real. This is proven by using symmetric polynomials and the following lemma.

Lemma.

$$\prod_{i=1}^n \prod_{j=1}^m (t - x_i y_j) = \det(tI - AB)$$

where A and B are matrices with entries expressed in terms of the elementary symmetric polynomials of x_i or y_j , respectively.

This lemma is interesting because it is easier to prove a more general statement which implies the lemma than it is to verify the lemma. We will prove the lemma when the x_i and y_j are independent variables, a more general statement than when the x_i and y_j represent numbers, but, nevertheless, an easier statement to prove.

Proof: Let

$$P_A(t) = \prod_{k=1}^n (t - x_k) = \sum_{l=0}^n (-1)^l \sigma_l(\mathbf{x}) t^{n-l}$$

and

$$P_B(t) = \prod_{j=1}^m (t - y_j) = \sum_{j=0}^m (-1)^j \sigma_j(\mathbf{y}) t^{m-j}.$$

Let

$$V_{k,l} = \begin{bmatrix} 1 \\ x_k \\ x_k^2 \\ \vdots \\ x_k^{n-1} \\ y_l \\ x_k y_l \\ \vdots \\ x_k^{n-1} y_l \\ \vdots \\ y_l^{m-1} \\ x_k y_l^{n-1} \\ \vdots \\ x_k^{n-1} y_l^{n-1} \end{bmatrix}$$

and let \bar{A} be the $n \times n$ matrix,

$$\bar{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ (-1)^{n+1} \sigma_n(x) & (-1)^n \sigma_{n-1}(x) & & & & \cdots & \sigma_1(x) \end{bmatrix}$$

Now let A be the following $nm \times nm$ matrix

$$A = \begin{bmatrix} \bar{A} & & & & \\ & \bar{A} & & & \\ & & \bar{A} & & \\ & & & \ddots & \\ & & & & \bar{A} \end{bmatrix}.$$

By plugging x_k into $P_A(t)$, we find that

$$x_k^n = \sum_{l=1}^n (-1)^{l+1} \sigma_l(x) x_k^{n-1}.$$

This implies that

$$AV_{k,l} = x_k \cdot V_{k,l}$$

where A is independent of k and l . In a similar way we can construct a matrix, B , with entries given by the elementary symmetric functions such that

$$BV_{k,l} = y_l V_{k,l}.$$

Now

$$\begin{aligned} ABV_{k,l} &= Ay_l V_{k,l} \\ &= y_l AV_{k,l} \\ &= x_k y_l V_{k,l}. \end{aligned}$$

Thus $\{x_k y_l\}$ are nm distinct roots of $\det(tI - AB)$ which is a monic polynomial of degree nm . Therefore,

$$\det(tI - AB) = \prod_{i=1}^n \prod_{j=1}^m (t - x_i y_j).$$

If the x_k 's and y_l 's were not independent variables, we would not be able to conclude that the elements in $\{x_k y_l\}$ are distinct.

With this lemma, we are ready to prove that the set of totally real numbers form a field under the operation $x \mapsto \sqrt{1+x^2}$.

Theorem. $\mathbb{F}_{\sqrt{1+x^2}} \subset \mathbb{F}_{TR}$.

Proof: If $\alpha, \beta \in \mathbb{F}_{TR}$, we must show that $-\alpha, \alpha^{-1}, \sqrt{1+\alpha^2}, \alpha + \beta, \alpha \cdot \beta \in \mathbb{F}_{TR}$. Let $\{\alpha_i\}_{i=1}^n$ be the conjugates of α and $\{\beta_j\}_{j=1}^m$ be the conjugates of β . We will prove the theorem by considering the following five polynomials.

$$\begin{aligned} q_{-\alpha}(t) &= \prod_{i=1}^n (t + \alpha_i), \\ q_{\alpha^{-1}}(t) &= \left(\prod_{i=1}^n (t - \alpha_i^{-1}) \right) \left(\prod_{i=1}^n \alpha_i \right), \\ q_{\sqrt{1+\alpha^2}}(t) &= \prod_{i=1}^n (t^2 - 1 - \alpha_i^2), \\ q_{\alpha+\beta}(t) &= \prod_{i=1}^n \prod_{j=1}^m (t - \alpha_i - \beta_j), \\ q_{\alpha\beta}(t) &= \prod_{i=1}^n \prod_{j=1}^m (t - \alpha_i \beta_j). \end{aligned}$$

The proofs of the first three cases are similar, and the proofs of the last two cases are similar, so we will only prove, in detail, the third case and the fifth case. If we expand $q_{\sqrt{1+\alpha^2}}(t)$, it is clear that the coefficients of t^k will be symmetric polynomials in the α_i . They may, therefore, be expressed as rational polynomials in the

elementary symmetric polynomials of the α_i . Since $(-1)^j \sigma_j(\alpha)$ are the coefficients of the minimal polynomial for α we may conclude that $q_{\sqrt{1+\alpha^2}}(t) \in \mathbb{Q}[t]$. It is clear that $\sqrt{1+\alpha^2}$ is a root of $q_{\sqrt{1+\alpha^2}}(t)$, thus the minimal polynomial of $\sqrt{1+\alpha^2}$, $p_{\sqrt{1+\alpha^2}}(t)$, divides $q_{\sqrt{1+\alpha^2}}(t)$. The fact that α is totally real implies that all of the conjugates, α_i , are real. Thus, $1+\alpha_i^2$ are all real and positive, so $\pm\sqrt{1+\alpha_i^2}$ are all real. We now conclude that all of the roots of $q_{\sqrt{1+\alpha^2}}(t)$ are real, and therefore $\sqrt{1+\alpha^2}$ is totally real.

For the fifth case, we use the previous lemma to conclude that $q_{\alpha\beta}(t) \in \mathbb{Q}[t]$. Clearly, $\alpha\beta$ is a root of $q_{\alpha\beta}(t)$ and all of the roots of $q_{\alpha\beta}(t)$ are real because α and β are totally real. In the other three cases, it is necessary to show that each of the q 's are polynomials with rational coefficients and only real roots. The first two cases may be tackled with the fact that the elementary symmetric polynomials generate the algebra of all symmetric polynomials. The fourth case may be verified with a lemma analogous to the previous lemma stating that

$$\prod_{i=1}^n \prod_{j=1}^m (t - x_i - y_j) = \det(tI - A - B).$$

This theorem gives us a practical way to decide that certain shapes may not be constructed using origami. For example, it is not possible using origami, to construct two cubes such that the volume of the second cube is twice that of the first cube. If this construction were possible, $\sqrt[3]{2}$ would be an origami number and would therefore be totally real. One, however, finds that the conjugates of $\sqrt[3]{2}$ are $\sqrt[3]{2}(-\frac{1}{2} \pm \frac{\sqrt{3}}{2}i)$ and $\sqrt[3]{2}$, but the first two are not real, so $\sqrt[3]{2}$ is not an origami number.

As we have seen before, $\sqrt{2} = \sqrt{1+1^2}$ and $\sqrt{4+2\sqrt{2}} = \sqrt{1+(1+\sqrt{2})^2}$ are origami numbers, so $\sqrt{2+\sqrt{2}} = \sqrt{2}^{-1}\sqrt{4+2\sqrt{2}}$ is an origami number. From this we see the following corollary.

Corollary. *It is not possible to construct a right triangle with arbitrarily given hypotenuse and leg using origami.*

Proof: If this were possible, it would be possible to construct a right triangle with hypotenuse $\sqrt{2+\sqrt{2}}$ and leg 1, since these are origami numbers. Any such triangle would have a leg of length $\sqrt{1+\sqrt{2}} = \sqrt{(\sqrt{2+\sqrt{2}})^2 - 1^2}$, but this is impossible because $\sqrt{1+\sqrt{2}}$ is not totally real.

The following corollary is a consequence of the standard algebraic description of compass and straight edge constructions and the two previous theorems [AH].

Corollary. *Every thing which is constructible with origami is constructible with a compass and straight edge, but the converse is not true.*

We want to expand on the relationship between compass and straight edge constructions and origami constructions. To review, compass and straight edge

constructions, let $\mathbb{F}_{\sqrt{x}}$ be the smallest subfield of \mathbb{C} closed under the operation $x \mapsto \sqrt{x}$, then $\mathbb{F}_{\sqrt{x}} \cap \mathbb{R}$ is the collection of numbers which are constructible with a compass and straight edge. From our work thus far, it is evident that the origami numbers, \mathbb{F}_0 , are contained in $\mathbb{F}_{\sqrt{x}} \cap \mathbb{F}_{TR}$. It is in fact the case that $\mathbb{F}_0 = \mathbb{F}_{\sqrt{x}} \cap \mathbb{F}_{TR}$. This characterization of the origami numbers is related to David Hilbert's 17th problem. At the International Congress of Mathematics at Paris in 1900, Hilbert gave a list of 23 problems [B]. His 17th problem was to show that any rational function which is non-negative when evaluated at any rational number is a sum of squares of rational functions. In 1926, Artin solved Hilbert's 17th problem [Ar]. The key idea which Artin used was the notion of totally positive. An element of a field is defined to be totally positive if it is positive in every order on the field. Artin proved that an element is totally positive if and only if it is a sum of squares. This is the idea which we use to prove the final characterization of the origami numbers.

Fact. [L, page 457]. If K is a finite real algebraic extension of \mathbb{Q} , then an element of K is a sum of squares in K if and only if all of its real conjugates are positive.

Theorem. $\mathbb{F}_0 = \mathbb{F}_{\sqrt{1+x^2}} = \mathbb{F}_{\sqrt{x}} \cap \mathbb{F}_{TR}$.

Proof: We have already shown that $\mathbb{F}_0 = \mathbb{F}_{\sqrt{1+x^2}}$ and that $\mathbb{F}_0 \subset \mathbb{F}_{\sqrt{x}} \cap \mathbb{F}_{TR}$, so we need to show that $\mathbb{F}_{\sqrt{x}} \cap \mathbb{F}_{TR} \subset \mathbb{F}_{\sqrt{1+x^2}}$. If $\alpha \in \mathbb{F}_{\sqrt{x}} \cap \mathbb{F}_{TR}$, then there exists a sequence of totally real numbers, $\{\beta_i\}_{i=1}^n$ and a sequence of totally real fields $\{K_j\}_{j=0}^{n-1}$ such that $K_0 = \mathbb{Q}$, $K_i = K_{i-1}(\beta_k)$, $\alpha = \beta_n$, and each β_i has degree 2 over K_{i-1} . Since β_i has degree 2 over K_{i-1} , β_i is a root of a polynomial of the form

$$x^2 + c_i x + d_i,$$

where $c_i, d_i \in K_{i-1}$. Therefore, $(\beta_i + c_i/2)^2 = c_i^2/4 - d_i$. By the proof of the previous theorem, we know that every conjugate of $(\beta_i + c_i/2)^2$ is the square of some conjugate of $\beta_i + c_i/2$. Hence, each of the conjugates of $(\beta_i + c_i/2)^2$ are positive and $(\beta_i + c_i/2)^2$ is a sum of squares of elements in K_{i-1} . Say that

$$(\beta_i + c_i/2)^2 = r_{i,1}^2 + r_{i,2}^2 + \cdots + r_{i,m}^2,$$

then,

$$\beta_i = r_{i,1} \sqrt{1 + \left[\frac{r_{i,2}}{r_{i,1}} \sqrt{1 + \left[\frac{r_{i,3}}{r_{i,2}} \sqrt{\cdots} \right]^2} \right]^2} - \frac{c_i}{2}$$

and we are done. This shows that any totally real number in $\mathbb{F}_{\sqrt{x}}$ is an origami number.

Legend has it that the ancient Athenians were faced with a plague. In order to remedy the situation, they sent a delegation to the oracle of Apollo at Delos. This delegation was told to double the volume of the cubical altar to Apollo. However, the Athenians doubled the length of each side of the altar, thereby creating an altar with eight times the volume rather than twice the volume of the original altar. Needless to say, the plague only got worse. For years, people have tried to double the size of a cube with compass and straight edge, and the gods have not smiled

upon them. We now can see that the gods will not be satisfied with our elementary origami either.

REFERENCES

[AH] G. Alexanderson, A. Hilman, *A First Undergraduate Course in Abstract Algebra*, 3rd edition, Wadsworth Publishing Company, 1983.
[Ar] E. Artin, *Über die zerlegung definiter funktionen in quadrate*, Abh. Math. Sem. Hausischen Univ. 5 (1927), 100–115.
[B] F. Browder, *Mathematical developments arising from Hilbert problems*, in “Proceedings of Symposia in Pure Mathematics,” Volume 28, American Mathematical Society, 1976.
[K] F. Klein, *Vortrage über ausgewahlte Fragen der Elementargeometrie*, Teubner, 1895.
[L] S. Lang, *Algebra*, 3rd edition, Addison-Wesley, 1993.
[M] J. Montroll, *Origami for the Enthusiast*, Dover, 1979.
[B] T. Sundara Row, *Geometric Exercises in Paper Folding*, Dover, 1966.

Department of Mathematics
The University of Texas at Austin
Austin, TX 78712

PICTURE PUZZLE
(from the collection of Paul Halmos)



This picture was taken in 1939.
(see page 242.)

An Abstract Algebra Story

Uri Leron and Ed Dubinsky

Statement: *The teaching of abstract algebra is a disaster, and this remains true almost independently of the quality of the lectures.*

We agree.

And we think there's a fairly wide consensus on this among experienced abstract algebra instructors, and an even wider one among experienced students.

Statement: *There's little the conscientious math professor can do about it. The stuff is simply too hard for most students. Students are not well-prepared and they are unwilling to make the effort to learn this very difficult material.*

We disagree.

But we suspect that many experienced abstract algebra instructors hold such beliefs. This is especially true for some excellent instructors: Their lectures are truly masterpieces, surely you can't improve much on *that*; so if the students still fail, that's too bad, but it can't really be helped.

We claim that, far from being an immutable fact of nature resulting from inadequacies of the student, this failure is, at least in part, an artifact of a too narrowly conceived view of instruction. In fact, *replacing the lecture method with constructive, interactive methods involving computer activities and cooperative learning, can change radically the amount of meaningful learning achieved by average students.*

In this paper we would like to paint a picture of such an alternative approach, which we and others have been developing and using in our classes over the last several years. We are painfully aware of the limitations inherent in any attempt to give such a description by means of the written text only. It would have been much better if you could actually visit our classes and observe the dynamics of the students' interactions with both the computer and their peers. By way of compromise, we will try to simulate such a visit by organizing our paper around several classroom "scenarios" and some commentary on the events depicted in each scenario. As a matter of principle, we have tried to make the scenarios as realistic as space limitation permits.

FIRST SCENARIO: WHAT IS IT ALL ABOUT?

Background. Our approach involves students working on computers in laboratory sessions and on their own. It also involves classroom discussions and assigned exercises, mainly with paper and pencil. The students work in teams which for the most part remain fixed for the entire semester.

This scenario takes place very early in the semester, after the students have done some computer work. In the exchange we are about to describe, the students are using a number of programs they have written in the programming language ISETL to implement the group axioms. The students wrote these programs after very brief and intentionally vague discussions (in the text and in class) of binary operations. The programs are quite simple and close to mathematics. For example, here is a program and one of its subprograms, expressing the definition of a group and existence of left-identity, respectively. The program `is_group` accepts as input a set G and a binary operation o and returns “true” or “false” according to whether the set with the binary operation is or is not a group. Similarly for `has_identity`.

```
is_group:= func(G,o);
    return is_closed(G,o)    and is_associative(G,o) and
           has_identity(G,o) and has_inverses(G,o);
end;
has_identity:= func(G,o);
    return exists e in G|(for all a in G|e.o a=a);
end;
```

Reality. You enter an abstract algebra class. The class is taking place in a microcomputer lab. The students are working on computer activities in teams of 2 to 4. The expressions they type at the keyboard, except for some minor details, look pretty much like standard mathematics; as a mathematician you have no trouble understanding the meaning of what they type, though you may be totally unfamiliar with computers. Before hitting the Return key, which would instruct the computer to evaluate their expression and exhibit the result, the students engage in a lively discussion trying to predict what this result is going to be. For example, you might observe something like the exchange in Figure 1¹.

Reflections. At this point you, the reader, have surely formed some hypotheses regarding our class. Let us now try to attend to some possible interpretations (and mis-interpretations) of what is really going on here. We present some likely hypotheses and questions posed by an idealized reader (**IR**), each followed by our reaction. In fact, these include some of the questions we have most often been asked when discussing our method with colleagues.

Idealized Reader (conjecturing): *The students are involved in a mathematical investigation concerning group-theoretic properties of modular arithmetic systems, also practicing at the same time their knowledge of the group axioms.*

True.

IR: *The students are learning by the “discovery method”.*

This is only partially true. In the first place, the students are indeed involved in posing questions and in trying to discover answers, but far from being left on their own to do that, they are guided by a worksheet provided by the instructor. For example, the activity in the above exchange might evolve in response to the

¹Real exchanges are characterized by a lengthy and messy zigzag path, and their transcription would take up more space than we can afford in this article. Therefore we are limited to bringing here only a stratified and compressed version of a real exchange. In fact we can include only brief, isolated fragments.

Figure 1. Student interactions within a mathematical computerized environment.

Note: the students' input is preceded by the '>' prompt; the computer's output is not.

<i>Interaction with the computer</i>	<i>Interaction within the team</i>
> Z12 := {0..11};	Why doesn't Z12 have 12 in it?
> a12 := x, y -> (x + y) mod 12 ;	What's the difference between Z12 and a12?
> is_group(Z12, a12);	What are the inputs to is_group?
true;	
> identity(Z12, a12);	What's happening here? I thought it would be true or false? What is <i>identity</i> supposed to return?
0;	What's wrong? Let's check the group properties.
> is_group(Z12-{0}, a12);	
false;	
> is_closed(Z12-{0}, a12);	Aha! So it's not closed, but I can't see why.
false;	Let's try some numbers.
> 7.a12 8 in Z12 - {0};	
true;	
> 7.a12 5 in Z12 - {0};	
false;	I know!
> a := x, y -> x + y ;	Let's try another operation.
> is_group(Z12, a);	
false;	
> 7.a 8 in Z12;	It's because $7 + 8 = 15$, not 3.
false;	
> m12 := x, y -> (x*y) mod 12 ;	What do you think about times?
> is_group(Z12, m12);	I think it will also be a group.
false;	Oops!
> is_closed(Z12, m12);	
true;	So it is closed...
> has_identity(Z12, m12);	... and it does have an identity...
true;	... but no inverses! What's an inverse, anyway?
> has_inverses(Z12, m12);	
false;	Oh! It must be zero. Let's take it out.
> is_invertible(Z12, m12, 0);	
false;	Oops! I don't understand it.
> is_group(Z12 - {0}, m12);	Not closed? But it was closed before... Must be the zero...
false;	What about this...
> is_closed(Z12 - {0}, m12);	
false;	
> 2.m12 6 = 0;	
true;	
> 2.m12 6 in Z12;	that's it. Let's try another mod? 11?
false;	Okay, this is the set...
> Z11 := {0..10};	... and the operation
> m11 := x, y := (x*y) mod 11 ;	now let's try it. I think it will work.
> is_group(Z11, m11);	
false;	Oops! Oh, yes, the zero again.
> is_group(Z11 - {0}, m11);	Now try it.
true;	Yeah!!!

following task:

Explore the modular systems Z_n (with or without 0) relative to addition and multiplication mod n . Formulate some conjectures, test them and try to give some explanations.

In the second place . . .

IR: *... But how can you expect students to discover in a few hours what took the best mathematical minds centuries?*

I was just getting there . . . In the second place, students are *not* expected to actually obtain complete, “correct” answers. The main purpose of the activities is to give them an *experiential basis* to which they can later relate the more abstract and formal treatments. Thus, subsequent discussions of a concept are more meaningful for students who have made a non-trivial effort trying to figure it out on their own, whether they have actually discovered it or not. Rather than “discovering” mathematical concepts, we think of our students as *constructing* them (in their mind). Computer experiences and classroom discussions are meant to help them make these mental constructions.

IR: *Doesn't learning how to program take a lot of time away from learning the mathematics?*

Because the syntax and basic constructs of ISETL are so close to standard mathematics, learning the language is inseparable from learning the mathematics--the programming “overhead” is minimal. You can check this for yourself by seeing if you have any trouble understanding the ISETL expressions in Figure 1. With our students, we spend only the first few sessions actually dealing with the language, and even then we do quite a bit of relevant math (like properties of the modular operations and of permutations). In writing their own code, and in using it to explore particular groups, students gain an understanding of the group concept of a quite different sort from that gained by listening to lectures and doing paper and pencil exercises (or by using the computer for drill and practice).

IR: *Still, wouldn't it be better if you gave them a software package in which functions like `is_group`, `has_identity`, `has_inverses` are pre-programmed? This way you'd save them the time of learning to program and the time of programming all these functions on their own, and still they would be able to “interact” with the computer.*

Programming `is_group` and the other functions is where the most important learning occurs. It is *the goal*, not just a *tool*. Experience, as well as modern learning theory (see, for example, Davis, Maher and Noddings, 1990 or Selden and Selden, 1990) tells us that one doesn't learn the group concept by memorizing the definition. In order to acquire meaning, the group concept has to be *constructed* in the learner's mind. Our method is based on the premise that if the students are asked to construct the group concept on the computer (by programming it), there is a good chance that a parallel construction will occur in their mind.

IR: *Are the computer activities meant to replace the traditional lecture?*

Yes. We think that the lecture method is, for most students, quite ineffective. Worse still, it makes them feel stupid, alienated. Most students tell us that they

have very little idea of what the lecture is about, even when it is delivered by a master lecturer. And if you try, say a year later, to see how much was retained from the course, you'll discover that it is close to nothing. (See, for example, Vinner, 1992.)

IR: *Are you assuming, then, that through the activities they get all the instruction they need?*

Not at all. The activities are followed by team work on assignments, team discussions, class discussions of subtle points, summary handouts (or assigned reading) of definitions and proofs, exercises etc. The role of the computer activities is, as we have said before, to provide an *experiential basis* for all the other learning modes. An important key to making this work is that students need to *reflect* on the computer (or paper and pencil) constructions that they make. A powerful stimulus for reflection is our insistence that they do their work in teams. Discussing and explaining what they are doing can lead students to make mental constructions that are parallel to the ones they are making on the computer.

When the students eventually encounter the “official”, general, abstract, formal version, this is perceived by them not as totally strange and prohibitive (as we believe is the case in standard lectures, where such abstractions are presented without any experiential basis), but as an elaboration of their previous experience. In popular terms we may say that the activities provide an initial intuitive familiarity with the topic to be learned. In more psychological terms (supported by an elaborate theoretical framework and research), we may say that the activities help the students to “construct” the mental processes, objects and relations necessary for a meaningful understanding of the topic.

IR: *I, too, believe in giving students intuitive explanations (such as the intuitive idea behind a complicated proof), which I always add to the formal part of my lectures. So why all the fuss?*

Experience, theory and research all point to the fact that verbal explanations that do not relate to the student's prior experience are quite ineffective (except for a few individuals with special talent in mathematics.) Intuition is the result of personal experience based on activity and interaction. A verbal representation of *your* intuitions (based on *your* past experience, activity and interaction) usually fails to re-create the same intuitions in the student's mind. Students need to construct the experience for themselves. Verbal explanations can help if they come *after* the student has formed an experiential basis. For then they serve to elaborate and conceptualize something the student has already vaguely known through the experience. We may say that verbal explanations can elaborate and explicate existing intuitions, but cannot *create* new ones.

IR: *Is your method so perfect? Surely there must be some difficulties, some unsolved problems, some things you are not so sure about . . .*

You must be kidding! The traditional method of teaching by lectures and exercises has developed over several hundred years. We feel that we have an approach which represents a significant improvement, but we are just beginning. Although we have been able to implement our method to obtain promising results in our own classes, and we are even beginning to learn how to disseminate this approach

to others, a multitude of problems remain and it will be a very long time before they are all solved.

There are difficulties inherent in the use of computers and with cooperative learning. It can be slow; it is not easy for a computer to deal with infinite objects; and the computer can only help indirectly with making proofs. Moreover, we cannot overemphasize the profound change in the teacher's attitude and the need to develop new skills that are part of adopting this method. Nevertheless, our theoretical perspective, experiences and research, and those of others, as well as reports from our students, all convince us that the improvement —the revolution! — in student learning justifies the effort.

SECOND SCENARIO: CONSTRUCTING LAGRANGE'S THEOREM AND ITS PROOF

Background. This topic is treated about one-third of the way into the semester. At that time the students, as well as the software environment, have grown considerably smarter. In particular, we describe three points which are necessary for understanding the next scenario.

1. At this point, the students are well-acquainted with groups, subgroups, and cosets, as well as with various examples, notably the modular groups Z_n (with addition mod n) and the symmetric groups S_n . All this knowledge is represented in ISETL by various functions and other mathematical objects which the students have constructed in previous activities. These are collected in a special initialization file (called *isettl.ini*) on the computer's disk (ISETL's long-term memory), and are automatically loaded each time ISETL is loaded. Thus *isettl.ini* can be considered to be a (dynamically changing) extension of the language, representing the collective, accumulative, official wisdom of the class.

2. The file *isettl.ini* at this point contains all the group and subgroup functions (is_closed, has_identity, is_group, is_subgroup, etc.) as well as some standard sets and operations such as Z_{12} , a_{12} (addition mod 12), S_4 (permutations of $\{1, 2, 3, 4\}$), os (permutation product), etc. It also contains the func PR, written previously by the students, which implements the definition of the "extended product" in a group. This program constructs and returns a func which accepts two inputs, decides if they are elements or subsets of the group and performs the appropriate operation between two elements, an element and a subset, or two subsets. Here is one version of it.

```
PR:= func(G,o);
  return func(x,y);
    if x in G and y in G then return x .o y;
    elseif x in G and y subset G then return { x.o b: b in y};
    elseif x subset G and y in G then return { a.o y: a in x};
    elseif x subset G and y subset G then return
      { a.o b: a in x, b in y};
  end; end; end;
```

Executing a statement such as $oo := PR(Z6, a6)$ will then make *oo* the name of the generalized product in this group.

3. Finally, the file *isettl.ini* contains an additional program called *name_group*, which assigns to a given group (and optionally a subgroup) all the standard notations. For example, assume that Z_{12} has been assigned in ISETL to denote the set of integers mod 12, a_{12} the operation of addition mod 12, and H_3 the

subgroup $\{0, 3, 6, 9\}$. If we now execute `name_group (Z12, a12, H3)`, then the names G, o, e, i, oo, H, K and $G \bmod H$ are automatically assigned the values, $Z12, a12, 0$, the inverse function, the extended product, $H3$, the coset function, and the set of all right cosets of $H3$ in G , respectively.

Students are encouraged to always use `name_group` in their computer investigations. The purpose is two-fold: First, using standard, short names facilitates the computer work. Secondly (and more profoundly), we believe that using “generic” names, and retaining the same names for different examples, helps students see the example under investigation as being a “generic” example; that is, it helps them in “seeing the general in the particular” (Mason & Pimm, 1984). This we believe, is a crucial step in the difficult and all-important processes of *generalization and abstraction*.

REALITY. The actual worksheet which the students receive appears in the left column of Figure 2. All of the mathematical expressions that are listed can be written mutatis mutandi in ISETL and evaluated on the computer. In the right column are a few remarks about what the students are expected to do and what we expect to be happening in their minds.

The entire activity represents about 60 minutes of cooperative lab work.

In Figure 2 we see students engaged in computer activities, prompted by a worksheet. Based on the experience gained in these activities, students are now further engaged in doing mathematical tasks and calculations in the classroom. Each task is allotted a certain amount of time and the students work on them cooperatively in their teams. While the students are working, the instructor has ample time to move around in the classroom, look at what the students are doing, answer questions and occasionally engage in a dialogue. The activity is followed by a class discussion. This may be the time for “official” summaries by the instructor. Some of the outcomes of the class activity are further pursued in homework assignments.

Figure 3 illustrates a typical classroom activity which follows the computer activities in the Lagrange’s theorem worksheet.

At this point, the instructor may decide that all of the ingredients of the proof of Lagrange’s theorem are present in the classroom and, more importantly, have been sufficiently constructed in the minds of most of the students. The complete formal proof may now be assigned as homework. Alternatively, or additionally, the instructor may tie everything together by presenting the formal statement of Lagrange’s theorem and its proof. There’s also a third alternative: Following the homework assignment, the instructor discusses with the students their work on the theorem and its proof, and then hands out for them a written page with the complete, polished proof on it (or refers them to the appropriate page in the course text).

Reflections. As in the first scenario, we now attend to some questions, observations and conjectures as might be posed by our “Idealized Reader”.

IR: *The func for PR looks to me like a hard program to write. Doesn’t it give students a lot of trouble, and doesn’t it involve too much programming effort?*

No on both counts. The main programming ingredient is the “if statement” with all the “elseif” clauses and by the time we get to this activity, the students are quite comfortable with such constructs. Other than this, the program is almost identical with the mathematical definition of this operation.

Figure 2. Tasks on cosets

Worksheet	Remarks
<p>H3 := {0, 3, 6, 9}, name_group(Z12, a12, H3); G; #G; 8 .o 9; is_group(G, o); e; i(7); H; is_subgroup(G, o, H);</p>	<p>Students are asked to predict the result of each expression, then enter it and resolve any conflicts between their predictions and the actual result. The goal for this first group of expressions is for students to refresh their memory about the meaning and interconnections of the many symbols involved in this activity.</p>
<p>H .oo 0; H .oo 1; H .oo 2; H .oo 3; H .oo 4; ... K(0); K(1); K(2); K(3); K(4); ...</p>	<p>Here students are moving on to familiarize themselves with cosets and their various notations. The operation K maps an element of G to its right H-coset. The "three dots" prompt the students to look for a pattern.</p>
<p>GmodH; #(GmodH); H subset G; H subset GmodH; H in GmodH; G in GmodH; K(1) in GmodH; K(2) in GmodH; {1, 4, 7, 10} in GmodH; {1, 4, 7, 10} subset GmodH; % union (GmodH) = G; forall a in G : a in K(a);</p>	<p>These are further prompts for exploring the set $G \text{ mod } H$ of all the right cosets of H in G, with a view to generality (again, using the predict-enter-resolve cycle). The operator "% union" is the extension of the binary operator "union" to operate on any (finite) set of sets.</p>
<p>Find as many equalities as you can among the sets $H, K(0), \dots, K(4)$, and among their cardinalities. Verify that your equalities return "true".</p>	<p>Computer exploration that could lead to one of the key steps in the proof of Lagrange's theorem.</p>
<p>When is it the case that the relation H .oo a = H .oo b is true?</p>	<p>Students are confronted with another issue in the proof.</p>
<p>Find examples of $i \neq j$ in G such that (one condition at a time): (a) $K(i) = K(j)$, (b) $K(i) \text{ inter } K(j) = \{ \}$, (c) neither (a) nor (b) holds.</p>	<p>The last major ingredient of the proof. (The symbol "\neq" is ISETL's way of approximating the mathematical "\neq" on the standard keyboard.)</p>
<p>What is the relation between the numbers #G, #H and #GmodH? Can you see an explanation?</p>	<p>Based on their previous activities and explorations, the students are now ready to try to discover Lagrange's theorem. Furthermore, since they have also discovered the properties of cosets which are responsible for the truth of the theorem, they now have a pretty good feeling for <i>why</i> the theorem is true.</p>
<p>Do name_group(Z12, a12, H4), where $H4 = \{0, 4, 8\}$, and predict the answers to all the previous activities. Check your predictions on the computer. Does the relation you found between #G, #H and #GmodH still hold?</p>	<p>If a student has begun to construct the idea of Lagrange's theorem and its proof, then looking at a second example may help bring it out.</p>
<p>Repeat the same activity with name_group(S3, os, A3).</p>	<p>More of the same with a non-commutative example.</p>

Figure 3. Worksheet on Lagrange's Theorem

Tasks	Outcomes
Formulate a conjecture regarding the orders of a finite group G , a subgroup H , and the set of cosets $G \bmod H$.	Some teams come up with the conjecture that the number of cosets times the order of the subgroup is equal to the order of the group. Fewer state that the order of the subgroup divides the order of the group. Some students do not see either of these explicitly, but have it on the tip of their tongues, so to speak: they recognize it readily upon hearing it from their peers.
Formulate these points in symbols.	Formulas such as $\#G = \#H * \#G \bmod H$ $o(G) = o(H) * \#G \bmod H$ $\#H \#G$ usually arise.
What are some properties of cosets which seem to "cause" these relations? Can you show how your conjecture follows from the properties you have listed?	The following are easy to observe from the previous activities, and some of them are commonly picked up by the students: — H is one of the cosets. — Every element belongs to some coset. — The union of all the cosets is the whole group. — Different cosets are disjoint. — If a and b belong to the same coset then $Ha = Hb$. — All the cosets have the same number of elements.

IR: *I notice that this program is supposed to create and return a new operation. I find that pretty confusing. Doesn't it give the students trouble?*

Yes, this is really difficult. This particular func takes as input a set and an operation (which is a function) and returns the generalized operation (which is also a function). One of the hardest things for students to do is to treat functions as total entities, or objects, and perform operations on them (Sfard, 1992). We see this in many situations, not only in algebra, but also in calculus and other subjects. Students expect a func to return a number or even a set of numbers and major mental adjustments are required before the student can understand that the func is to return a "whole function".

But this difficulty is with the mathematics, not the programming. We have done studies which suggest that confronting this kind of mathematical difficulty in a computer context can help students develop the ability to work with functions as objects (Ayres, Davis, Dubinsky, and Lewin, 1986, Breidenbach, Dubinsky, Hawks, and Nichols, 1991).

IR: *Some of the expressions your students write are more formal than in the usual abstract algebra classroom. In my experience, students find such formalities quite hard and bizarre.*

This had been our experience too—as long as we were trying to *present* such formalities in a lecture. We were surprised and delighted to see students arriving

at such subtleties on their own, aided by feedback from the computer and the on-going discussions. The ISETL medium makes it both necessary and possible for students to deal with such subtleties. The need to “explain” a mathematical concept to a dumb computer is a wonderful motivation for using formal language; watching how the beast “understands” (or mis-understands) your explanations, is a great way to climb to the next step in the ladder of successive refinements.

We would like to suggest that there is a distinction between *formalism* and *empty formalism*. When students are asked to deal with a formal statement *before* they have an opportunity to construct the processes and objects that the formality describes, then it is empty formalism. When the formal statement is a description of ideas which already exist in the student’s mind, then the symbolism becomes a convenient way of communicating these ideas and can even be a powerful tool for further mathematical growth.

IR: *Asking the students (in your worksheet) to characterize the relation “ $Ha = Hb$ ” seems to me like a regular pencil-and-paper exercise.*

This is true, and if a student prefers to do it this way, that’s fine with us. However, many students merely get stuck when facing such a question with only pencil-and-paper as working tools. In the ISETL lab, students still make errors but they rarely get stuck, since they can utilize the computer to analyze examples and conduct investigations. In our experience, most students prefer an environment in which they can utilize such “experimental scaffolding” to aid the theoretical reasoning.

IR: *Do the students really discover all of those coset properties on their own?*

Some do and some don’t. It is wonderful for those who do, but again, the point is that this is not necessary. The effort that students put into *trying* to discover these facts is sufficient to change the way in which they respond internally when they hear it from their classmates, or from the instructor.

IR: *Isn’t it frustrating for the students who fail to accomplish the given tasks?*

It is very important for the instructor to make sure students don’t conceive of this as *failure*. In our classes, we explicitly discuss this issue with the students, and emphasize many times that the main point is spending the time and effort on the problem, not solving it.

IR: *Well, you have said a lot about alternatives to lecturing, but here, in the end, what you are doing with Lagrange’s theorem is to give a lecture about the theorem and its proof.*

We have no objection to lecturing as such. Our main claim is that introducing new material via a lecture may be a very effective *teaching* method, but it is mostly a very ineffective *learning* method. We use (short!) lectures to summarize and elaborate something the students have previously spent considerable time and effort working on. Ideally, the lecture is on something the students are already familiar with; it only serves to present it in a more explicit, general, precise, formal way.

IR: *The activities have directed the students to some ingredients of the proof, but many remain untouched. For example, they may have discovered that different cosets are disjoint, but there’s nothing in the activities to help them actually prove this.*

True. In a way we may say that the students now see the proof as being based on a few “main ideas” of the sort you have mentioned, but they haven’t yet examined the proof of these main ideas themselves. This “layered” view of the proof is called a *structural proof* and has actually some advantages (Leron, 1983, 1985). In this sense we might say that the students are now familiar with the “top level” of the proof, but they still need to supply the details of the lower levels. In practice, these missing details, are usually assigned as homework.

THIRD SCENARIO: A GENTLE INTRODUCTION TO NORMALITY AND QUOTIENT GROUPS

Background. The situation at this point in the course is similar to that described at the beginning of the second scenario, and we are making similar use of the file *isetl.ini* and the procedure *name_group* discussed there. At this point the students have at their disposal, and are pretty comfortable with, the set $G \bmod H$ of all right cosets of a subgroup H in G , and the “extended product” operation oo defined on it. This operation had already been extensively used to construct and investigate cosets, but at this point, they have hardly used the fact that *the same operation can also be used to multiply two cosets*.

Note: We have chosen here a somewhat non-standard route to approaching the notions of normality and the quotient group. In particular, the set $G \bmod H$ used in the ISETL activities is defined whether H is normal or not. Also, we are initially making an unusual choice for our definition of coset product. Eventually, for a *normal* subgroup H in G , $G \bmod H$ and this product coincide with the usual G/H and the usual definition of product “by representatives”. But with our approach they can be used to great advantage before normality has even been introduced.

The activity we are about to discuss takes place a little before we are half-way through the course. It typically starts with a short (10–15 minutes) class discussion, in which the previous activities and results concerning subgroups and their cosets are recalled.

It is then noted that, since oo enables us to multiply any two cosets, the question naturally arises as to whether (or, better yet, *when*) $G \bmod H$ is a group under this operation. Next, the students do some “warm-up” activities on the computer to remind them of various definitions and concepts. The students then proceed to do a hands-on investigation, guided by worksheets as shown in Figures 4, 5.

Reality—Normality and Quotient group worksheet.

Note: As a space-saving device, we are writing several expressions to a line; in reality the students enter them one by one.

The computer activities shown in Figures 4 and 5 last for about 50 minutes. The students are given instructions which prompt them to focus their attention on the relevant aspects of a complex situation. One of the pleasant features of this kind of non-prescriptive learning environment is that those who can make sense of such advice will, while those who can’t, will just ignore it (at least for the moment), with no harmful side-effects.

Here are the instructions given to the students.

Figure 4. Products of cosets

Following the warm-up activities the students start to look into the product of cosets and its properties. Some of these expressions are saying the same thing, but with different notation and on different levels. It is important (and non-trivial) for students to be able to experience these different levels of expression as being “the same”. As for the main question, at this point many tend to over-generalize, believing that $G\text{mod}H$ is always a group.

```
> name_group(Z12, a12, {0, 3, 6, 9});
Group objects defined:      G, o, oo, e, i.
Subgroup objects defined:   H, GmodH, K.
> K(1) .oo K(2);
{0, 3, 6, 9};
> K(1) .oo K(2) in GmodH;
true;
> K(1) .oo K(2) = K(0);    K(0) .oo K(1) = K(1);    K(2) .oo K(4) = K(9);
true; true; false;
> for all a, b in G|K(a) .oo K(b) = K(a + b);
false;
> for all a, b in G|K(a) .oo K(b) = K(a .o b);
true
> for all a in G|H .oo a = a .oo H;
true;
> exists x, y in GmodH|x .oo y notin GmodH;
false;
> is_closed(GmodH, oo);
true;
> is_group(GmodH, oo);
true;
> identity(GmodH, oo);
{0, 3, 6, 9};
```

In the following activities, predict the answer and then check on the computer. Try to settle any discrepancies that arise. Please pay special attention to the following points and write down any observations you may come up with.

- *When is the product of two cosets again a coset?*
- *When is $(Ha)(Hb) = H(ab)$?*
- *When is $G\text{mod}H$ closed under the operation oo ?*
- *When is $G\text{mod}H$ a group?*

The last part of the classroom session lasts about 20 minutes, and will not be described here in detail. The instructor leads an interactive discussion in which the different teams share their discoveries, conjectures and questions. By building on the material that the students bring up, the instructor is able to state most naturally and smoothly the definition of a normal subgroup, the theorem that when H is normal then $G\text{mod}H$ forms a group, and the (now very easy) proof of this theorem. Normality is naturally introduced here as the condition which insures that $G\text{mod}H$ be a group, and the definition most often discovered by the students is $aH = Ha$ for all $a \in G$. Except for the new name, the students can really feel that the instructor merely summarizes what they have found in their investigations. In the session that follows, the instructor makes the final ties with the “standard” approach by explaining that when H is normal, $G\text{mod}H$ is commonly denoted G/H , and is called the *quotient group of G modulo H* and coset product is commonly *defined* by the formula $(Ha)(Hb) = H(ab)$. In our approach, this formula receives the status of a theorem that comes up in the activities.

Figure 5. A more complicated case

The second round through the activities, which we omit, repeats the experience of Figure 4 with a non-normal example. At this stage, some students over-react and decide that $G\text{mod}H$ is a group if and only if G is commutative. Coming to the most “mature” example so far (a normal subgroup in a non-commutative group), students start to realize that they need to look deeper into the properties of H that make $G\text{mod}H$ a group. They also come to appreciate that the main issue here is closure, namely, when is the product of two cosets again a coset. More specifically, when is it the case that $HaHb = Hab$? This leads rather smoothly and naturally to the desired property $Ha = aH$, which is one way of defining normality. It is very important to note that the way they have “constructed” it for themselves, normality is immediately perceived as a way to ensure that $G\text{mod}H$ is a group.

```
> A3 := {[1, 2, 3], [2, 3, 1], [3, 1, 2]};
> name_group(S3, os, A3);
Group objects defined: G, o, oo, e, i.
Subgroup objects defined: H, GmodH, K.
> G; #G;
{[1, 2, 3], [1, 3, 2], [2, 1, 3], [2, 3, 1], [3, 1, 2], [3, 2, 1]}; 6;
> H; #H;
{[1, 2, 3], [2, 3, 1], [3, 1, 2]}; 3;
> GmodH; #GmodH;
{([1, 2, 3], [2, 3, 1], [3, 1, 2]), ([1, 3, 2], [2, 1, 3], [3, 2, 1])}; 2;
> forall p, q in G[K(p) .oo K(q) = K(p .o q);
true;
> forall p in G[H. oo p = p .oo H;
true;
> is_group(GmodH, oo);
true;
> identity(GmodH, oo);
{[1, 2, 3], [2, 3, 1], [3, 1, 2]};
```

Reflections.

Idealized Reader: *In my experience, the quotient group is one place in the course where most students experience a real crisis of meaning. They just stare at the symbols, or hear the instructor’s words, and all they see (or hear) is so many ink stains (or so many words). Is there a difference in your class?*

There’s no question that this is a difficult concept, and in our class too students have to struggle hard. However, because of the computer activities and the accompanying discussions in the teams, they now have a way of constructing meaning by “successive refinement”, as they keep updating and refining their understanding by watching and trying to understand the computer’s response to the mathematical expressions they enter.

IR: *What do you mean by “constructing meaning”?*

We believe that this issue is at the heart of students’ difficulties in abstract algebra and mathematics in general. Students often have great difficulty with a theorem and its proof, such as the Homomorphism theorem or even Lagrange’s theorem. We tend to explain this difficulty by saying “the theorem is complicated”. But is it really the *theorem* that is so complicated? The theorem and its proof are, essentially, about certain relationships which hold between certain mathematical objects. These relationships are fairly simple and we believe that the *difficulty*

experienced by the students lies not so much in the complexity of the theorem, as in the abstract nature of the mathematical objects involved.

IR: *Can you give an example?*

Yes, let's look at the relationships involved in Lagrange's theorem and its proof: one number dividing the other, two sets having the same cardinality or being disjoint, etc. These are easy enough. The *objects*, on the other hand, are "complicated" in the sense of the many levels of abstraction and the great time and effort needed to "construct" them in the mind of the student. For us, the simplicity of the proof lies in our ability to have a clear image of the group as being partitioned into a disjoint union of cosets. But in order for the students to have such an image, they need to "construct" in their mind not only "group", "sub-group" and "coset" but also "the set of all cosets of H in G ".

One powerful aspect of the computer environment we have been describing is that constructing these new entities on the computer helps students see them as real things that have their own properties, and that can be manipulated, investigated and discussed.

IR: *How can you tell that your method really makes a difference?*

The fact is that students in our classes don't feel the same alienation and paralysis in the face of the quotient group, and end up with a good understanding of it. This is borne out in the research that accompanies our teaching, both in listening to the students' own subjective evaluation of the method and their learning in it, and by in-depth interviews of many individual students. At the very least, they have little difficulty, even on exams, in constructing specific quotients, for example in S_4 , and identifying the resulting group as being isomorphic to some known groups. Our data suggests that this understanding tends to carry over into ring theory where they can learn to construct and analyze quotients of polynomial rings by various ideals.

IR: *How can you explain the improvement in students' understanding?*

We think it is due to our method which is based on modern understanding of how people learn. This involves a dramatic change in the basic assumptions about how learning occurs: More and more people in our profession are beginning to embrace the notion that knowledge is not *transferred* from one person (the instructor) to another (the student), but rather is *constructed* in the learner's mind. This mental construction takes time and effort and requires an appropriate learning environment the design of which is informed by research into how mathematics is learned.

IR: *Exactly how do you use research here?*

The role of research in this context is to propose specific mental constructions that students can make in order to learn specific mathematical concepts, and to propose and evaluate methods for fostering such constructions. For example, to go into a little more detail on a point that was discussed earlier, our research suggests that students' difficulty with understanding Lagrange's theorem may be largely due to their confusion about the nature of cosets. We find that they can understand the *process* of forming a coset, but often cannot take the next step of seeing these cosets as *objects* to be measured, counted and compared. Thus, we design computer activities aimed at getting students to construct cosets on the computer

and then treat them as objects by manipulating them in various ways. We have done this for a number of mathematical concepts and this work forms a major component of our efforts. We are beginning to develop a body of literature about it which the reader can consult (Ayres et al. 1986; Breidenbach et al, 1991; Dubinsky, 1986, 1989; Dubinsky, Dautermann, Leron and Zazkis, 1994; Dubinsky, Elterman and Gong, 1989; Dubinsky and Leron, 1994).

CONCLUSION. In this paper we have attempted to describe a new paradigm for teaching undergraduate mathematics in general and abstract algebra in particular. We have started out from what we perceive as a general discontent by instructors and students alike with the state of teaching abstract algebra, and have marshaled all the resources at our disposal to develop this new paradigm. These resources include contemporary theory and research into the mental processes involved in learning mathematics, collaborative learning methods, the use of computers equipped with a custom-made programming language, and our own experience over many years of teaching undergraduate mathematics.

These methods are not restricted to Abstract Algebra but have also been applied to courses in Discrete Mathematics, Precalculus and Calculus.

We have presented here a “constructivist” approach, according to which *telling* students about mathematical processes, objects and relations is not sufficient to induce meaningful learning (hence the sorry state of affairs even with the best of lecturers). What is required, rather, is a learning environment which encourages and enables students to make *mental constructions* corresponding to these mathematical processes, objects and relations. Forming such learning environments is a highly non-trivial educational task and, in our opinion, the best way of inducing such mental constructions is by having students make appropriate constructions on a computer, and by using the social context to reflect on the computer activities. Our main tool for making constructions on the computer has been programming in ISETL, while our main tool for reflecting on the activities has been collaborative learning. Along with the computer activities and the accompanying discussions in teams, we of course also use the standard tools of class discussion, short summaries by the instructor and homework assignments. However, it is our belief that these tools are most effective when applied *after* the activities.

Making the transition to teaching with this constructivist, interactive method is not easy. In addition to making fundamental changes in long-held attitudes about teaching and learning, the instructor must make a substantial initial investment of time and energy for learning techniques very different from the standard lecture method. But our experience and research is showing that this extra effort on the part of the instructor is extremely well rewarded by the change in students’ attitudes towards the course and mathematics in general, and by the amount of meaningful learning which is achieved.

REFERENCES

-
- Ayres, T., G. Davis, E. Dubinsky, and P. Lewin, (1986), Computer experiences in learning composition of functions, *Journal for Research in Mathematics Education*, 19, 3, 246–259.
- Breidenbach, D., E. Dubinsky, J. Hawks, and D. Nichols, (1991), Development of the process concept of function, *Educational Studies in Mathematics*, 247–285.
- Clement, J., J. Lochhead, and E. Soloway, (1980), Positive effects of computer programming on students’ understanding of variables and equations, *Comm. ACM*. 467–474.
- Davis, R. B., C. A. Maher, and N. Noddings, (1990), Constructivist Views on the Teaching and Learning of Mathematics, *Journal for Research in Mathematics Education*, Monograph, No. 4, Reston: NCTM.

- Dubinsky, E., (1986), Teaching mathematical induction I, *The Journal of Mathematical Behavior*, 5, 305–317.
- Dubinsky, E., (1989), Teaching mathematical induction II, *The Journal of Mathematical Behavior*, 8, 285–304.
- Dubinsky, E., (in press), On learning quantification, in M. S. Arora. (ed.), *Mathematics Education: The Present State of the Art*. UNESCO.
- Dubinsky, E., J. Dautermann, U. Leron, and R. Zazkis, (1994), On learning fundamental concepts of group theory, *Educational Studies in Mathematics*, 27, 267–305.
- Dubinsky, E., F. Elterman, and C. Gong, (1989), The student's construction of quantification, *For the Learning of Mathematics* 8, 2, 44–51.
- Dubinsky, E. and U. Leron, (1994), *Learning Abstract Algebra with ISETL*. New York: Springer.
- Dubinsky, E. and K. Schwingendorf, (1992), *Calculus, Concepts, and Computers: Preliminary version*, St. Paul: West.
- Leron, U., (1983), Structuring mathematical proofs, *American Mathematical Monthly* 90, 174–185.
- Leron, U., (1985), Heuristic presentations: The role of structuring, *For the Learning of Mathematics*, 5, 3, 7–13.
- Mason, J. and D. Pimm, (1984), Generic examples: Seeing the general in the particular, *Educational Studies in Mathematics*, 15 (3), 277–289.
- Papert, S., (1980), *Mindstorms: Children, Computers and powerful ideas*, Basic Books.
- Selden, A. and J. Selden, (1990), Constructivism in mathematics education: A view of how people learn, *UME Trends*, 2, 1, p. 8.
- Sfard, A., (1992), Operational origins of mathematical notions and the quandary of reification—the case of function. In G. Harel and E. Dubinsky (Eds.), *The Concept of Function: Aspects of Epistemology and Pedagogy*. MAA Notes Series No. 25, Math Assn. Amer.
- Vinner, S., (1992), The function concept as a prototype for problems in mathematics learning In G. Harel and E. Dubinsky (Eds.), *The Concept of Function: Aspects of Epistemology and Pedagogy*. MAA Notes Series No. 25, Math. Assn. Amer.

Department of Science Education
Technion-Israel Institutes of Technology
Haifa 32000, ISRAEL
uril@techurix.technion

Mathematics Department
Purdue University
West Lafayette, IN 47907
bbf@sage.cc.purdue.edu

“I think you’re begging the question,” said Haydock, “and I can see looming ahead one of those terrible exercises in probability where six men have white hats and six men have black hats and you have to work it out by mathematics how likely it is that the hats will get mixed up and in what proportion. If you start thinking about things like that, you would go round the bend. Let me assure you of that!”

—Agatha Christie

The Mirror Crack’d. Toronoto: Bantam Books, 1962.

Answer to Picture Puzzle
(p. 226)
André Weil

A Multidimensional Version of Rolle's Theorem

Massimo Furi and Mario Martelli

In this paper we obtain for functions $f: \mathbf{R}^n \rightarrow \mathbf{R}^p$ a version of Rolle's Theorem which we hope the readers will find useful and interesting for the following reasons. Three fundamental results from Calculus: namely Rolle's Theorem, the Mean Value Theorem and the Cauchy Generalized Mean Value Theorem can be easily derived from it. The version has intuitive geometrical applications and the proof is very simple.

Teachers may find it appropriate to incorporate our result in a course on Multivariable Calculus, since it provides an example of how certain one-dimensional theorems can be rephrased in higher dimensional spaces, and it shows that by expanding our mathematical horizon we frequently gain in organization and unity. Professional mathematicians are all familiar with these facts, but students will surely derive from them a motivation to learn more.

The basic idea of our result is to assume a certain behavior of f on the boundary ∂R of a n -dimensional region R (in the real line this behavior reduces to the familiar condition $f(a) = f(b)$) to obtain information on the derivative of f at an interior point of R . Of particular relevance to the result is the Mean Value Theorem of Sanderson [10] for a function $v: [a, b] \rightarrow \mathbf{R}^p$. We extend his theorem to functions of several variables.

The paper ends with an additional, more general version of Rolle's Theorem, and with an open problem and a conjecture which will hopefully stimulate the reader's mathematical curiosity.

We now list the terminology used and the results needed in the sequel. $\mathbf{0}(m \times n)$ stands for the zero matrix with m rows and n columns. $\mathbf{x} \cdot \mathbf{y}$ denotes the Euclidean inner product between \mathbf{x} and \mathbf{y} and the norm of \mathbf{x} is $\|\mathbf{x}\| = \sqrt{\mathbf{x} \cdot \mathbf{x}}$. We repeatedly make reference to the following sets:

$$D(\mathbf{x}_0, r) = \{\mathbf{x} \in \mathbf{R}^n: \|\mathbf{x} - \mathbf{x}_0\| \leq r\}, B(\mathbf{x}_0, r) = \{\mathbf{x} \in \mathbf{R}^n: \|\mathbf{x} - \mathbf{x}_0\| < r\},$$

$$\text{and } S(\mathbf{x}_0, r) = \{\mathbf{x} \in \mathbf{R}^n: \|\mathbf{x} - \mathbf{x}_0\| = r\} = \partial D(\mathbf{x}_0, r).$$

The two propositions below play a key role in the proof of our multi-dimensional version of Rolle's Theorem.

Proposition 1. *Let $f: D(\mathbf{x}_0, r) \subset \mathbf{R}^n \rightarrow \mathbf{R}$ and let $\mathbf{c} \in B(\mathbf{x}_0, r)$ be an extremum point of f . Assume that f is differentiable at \mathbf{c} . Then $f'(\mathbf{c}) = \mathbf{0}(1 \times n)$.*

Proposition 2. *Let $f: D(\mathbf{x}_0, r) \subset \mathbf{R}^n \rightarrow \mathbf{R}$ be continuous. Then the image of f is a closed and bounded interval $[m, M]$.*

We point out that the proof of Rolle's Theorem in \mathbf{R} is based on the one-dimensional version of the two propositions.

Results. The following simple example shows that a straightforward reformulation of Rolle's Theorem in \mathbf{R}^n , $n \geq 2$, fails.

Example 1. Let $f: \mathbf{R}^2 \rightarrow \mathbf{R}^2$ be defined by

$$f(x, y) = (x(x^2 + y^2 - 1), y(x^2 + y^2 - 1)).$$

The function f is continuous on $D(\mathbf{0}, 1)$, is differentiable on $B(\mathbf{0}, 1)$ and $f(\mathbf{x}) = \mathbf{0}$ for every $\mathbf{x} \in S(\mathbf{0}, 1)$. However, $f'(\mathbf{x}) \neq \mathbf{0}(2 \times 2)$ for all $\mathbf{x} \in B(\mathbf{0}, 1)$.

We are now ready to state and prove our main result.

Theorem 1. Let $f: D(\mathbf{x}_0, r) \subset \mathbf{R}^n \rightarrow \mathbf{R}^p$ be continuous on $D(\mathbf{x}_0, r)$ and differentiable on $B(\mathbf{x}_0, r)$. Assume that there exists a vector $\mathbf{v} \in \mathbf{R}^p$ such that

$$\text{i) } \mathbf{v} \text{ is orthogonal to } f(\mathbf{x}) \text{ for every } \mathbf{x} \in S(\mathbf{x}_0, r).$$

Then there exists a vector $\mathbf{c} \in B(\mathbf{x}_0, r)$ such that $\mathbf{v} \cdot f'(\mathbf{c})\mathbf{u} = 0$ for every $\mathbf{u} \in \mathbf{R}^n$.

Proof: Let $k: \mathbf{R}^p \rightarrow \mathbf{R}$ be defined by $k(\mathbf{x}) = \mathbf{v} \cdot \mathbf{x}$. Set $g(\mathbf{x}) = k(f(\mathbf{x}))$. By Proposition 2 the image of g is a bounded and closed interval $[m, M]$. Assumption i) implies that g is 0 on $S(\mathbf{x}_0, r)$. Hence we may assume, without loss of generality, that g reaches its maximum value, M , at a point $\mathbf{c} \in B(\mathbf{x}_0, r)$, namely $M = g(\mathbf{c})$. By Proposition 1 $g'(\mathbf{c}) = \mathbf{0}(1 \times n)$, i.e. $\mathbf{v} \cdot f'(\mathbf{c})\mathbf{u} = 0$ for every $\mathbf{u} \in \mathbf{R}^n$. QED.

Remark 1. Assumption i) can be replaced by the equivalent statement

$$\text{"ii) } \mathbf{v} \cdot f(\mathbf{x}) \text{ is constant on } S(\mathbf{x}_0, r)\text{"};$$

and the conclusion of the theorem can be expressed in the equivalent but geometrically more intuitive way

$$\text{"}\mathbf{v} \text{ is orthogonal to the vectors } \frac{\partial f}{\partial x_1}(\mathbf{c}), \frac{\partial f}{\partial x_2}(\mathbf{c}), \dots, \frac{\partial f}{\partial x_n}(\mathbf{c})\text{"}.$$

Remark 2. $D(\mathbf{x}_0, r)$ can be replaced by the closure of any open, bounded and connected set R of \mathbf{R}^n .

Rolle's Theorem, the Mean Value Theorem and the Cauchy Generalized Mean Value Theorem are easily derived from Theorem 1.

Corollary 1 (Cauchy). Let $a < b$ and $f, g: [a, b] \rightarrow \mathbf{R}$ be continuous on $[a, b]$ and differentiable on (a, b) . Then there exists $c \in (a, b)$ such that

$$[f(b) - f(a)]g'(c) = f'(c)[g(b) - g(a)].$$

Proof: If $f(a) = f(b)$ and $g(a) = g(b)$ there is nothing to prove.

Assume $[f(b) - f(a)]^2 + [g(b) - g(a)]^2 > 0$. Define $S: [a, b] \rightarrow \mathbf{R}^2$ by $S(t) = (g(t), f(t))$. Let $\mathbf{v} = (f(b) - f(a), g(b) - g(a))$. Then $\mathbf{v} \cdot T(a) = \mathbf{v} \cdot T(b) = f(b)g(a) - f(a)g(b)$. Hence, according to Theorem 1 (see Remark 1), there is a point $c \in (a, b)$ such that $\mathbf{v} \cdot T'(c)t = 0$ for every $t \in \mathbf{R}$. With $t \neq 0$ we obtain $[f(b) - f(a)]g'(c) = f'(c)[g(b) - g(a)]$.

Setting $g(x) = x$ gives the Mean Value Theorem. If, in addition $f(b) = f(a)$, then we have Rolle's Theorem. QED

The next Corollary is the Mean Value Theorem of Sanderson [10] mentioned in the Introduction.

Corollary 2. *Let $a < b$ and $\mathbf{v}: [a, b] \rightarrow \mathbf{R}^p$ be k times differentiable. Assume that $\mathbf{v}(a), \mathbf{v}(b)$ and the $k - 1$ derivatives of \mathbf{v} at a are orthogonal to a non-zero vector \mathbf{v}_0 . Then, for some $c \in (a, b)$, $\mathbf{v}^{(k)}(c)$ is orthogonal to \mathbf{v}_0 .*

Proof: From Theorem 1 we derive the existence of a point $c_1 \in (a, b)$ such that \mathbf{v}_0 is orthogonal to $\mathbf{v}'(c_1)$. The theorem can now be applied to \mathbf{v}' in the interval $[a, c_1]$ to yield a point $c_2 < c_1$ such that \mathbf{v}_0 is orthogonal to $\mathbf{v}''(c_2)$. This procedure can be repeated $k - 1$ times to obtain $c = c_k < c_{k-1}$ such that $\mathbf{v}_0 \cdot \mathbf{v}^{(k)}(c) = 0$. QED

A recent result of Evard and Jafari [4] (see also [7]) follows from Theorem 1.

Corollary 3. *Let \mathbf{C} be the field of complex numbers and $f: \mathbf{C} \rightarrow \mathbf{C}$ be a holomorphic function. Assume that there are points $\mathbf{a} \neq \mathbf{b}$ such that $f(\mathbf{a}) = f(\mathbf{b})$. Then there exist $\mathbf{z}_1, \mathbf{z}_2$ in the open line segment joining \mathbf{a} with \mathbf{b} such that $\text{Re}(f'(\mathbf{z}_1)) = \text{Im}(f'(\mathbf{z}_2)) = 0$.*

Proof: Let $f(\mathbf{z}) = f(x + iy) = u(x, y) + iv(x, y)$ and $\mathbf{p} \in \mathbf{R}^2$, $\mathbf{p} = (p_1, p_2) = (\text{Re}(\mathbf{a}), \text{Im}(\mathbf{a}))$, $\mathbf{q} \in \mathbf{R}^2$, $\mathbf{q} = (q_1, q_2) = (\text{Re}(\mathbf{b}), \text{Im}(\mathbf{b}))$. Define $g(t) = (u(\mathbf{q} + t(\mathbf{p} - \mathbf{q})), v(\mathbf{q} + t(\mathbf{p} - \mathbf{q})))$, $t \in [0, 1]$. Notice that $g(0) = g(1)$. According to Theorem 1, for every $\mathbf{x} \in \mathbf{R}^2$, $\mathbf{x} \neq \mathbf{0}$, there exists $t_0 \in (0, 1)$ such that $\mathbf{x} \cdot g'(t_0)t = 0$, for every $t \in \mathbf{R}$. Let $t = 1$ and choose the vector $\mathbf{x}_1 = (p_1 - q_1, p_2 - q_2)$. Then

$$\begin{aligned} 0 = \mathbf{x}_1 \cdot g'(t_0) &= \frac{\partial u}{\partial x}(g(t_0))(p_1 - q_1)^2 + \frac{\partial u}{\partial y}(g(t_0))(p_1 - q_1)(p_2 - q_2) \\ &\quad + \frac{\partial v}{\partial x}(g(t_0))(p_1 - q_1)(p_2 - q_2) + \frac{\partial v}{\partial y}(g(t_0))(p_2 - q_2)^2. \end{aligned}$$

Since f is holomorphic, its real and imaginary part satisfy the Cauchy-Riemann equations (see [1]), i.e. $\partial u / \partial x = \partial v / \partial y$ and $\partial u / \partial y = -\partial v / \partial x$. Hence

$$\frac{\partial u}{\partial x}(g(t_0))[(p_1 - q_1)^2 + (p_2 - q_2)^2] = 0.$$

This implies $\partial u / \partial x(g(t_0)) = \text{Re}(f'(\mathbf{z}_1)) = 0$, where $\mathbf{z}_1 = \mathbf{q} + t_0(\mathbf{p} - \mathbf{q})$.

To obtain the other equality use the vector $\mathbf{x}_2 = (q_2 - p_2, p_1 - q_1)$. QED

Theorem 1 can be given a slightly more general form.

Theorem 2. (Second version of Rolle's Theorem in \mathbf{R}^n). *Let $f: D(\mathbf{x}_0, r) \subset \mathbf{R}^n \rightarrow \mathbf{R}^p$ be continuous on $D(\mathbf{x}_0, r)$ and differentiable on $B(\mathbf{x}_0, r)$. Let $\mathbf{v} \in \mathbf{R}^p$, $\mathbf{z}_0 \in B(\mathbf{x}_0, r)$ be such that*

$$\text{ii) } \mathbf{v} \cdot (f(\mathbf{x}) - f(\mathbf{z}_0)) \text{ does not change sign on } S(\mathbf{x}_0, r).$$

Then there exists a vector $\mathbf{c} \in B(\mathbf{x}_0, r)$ such that $\mathbf{v} \cdot f'(\mathbf{c})\mathbf{u} = 0$ for every $\mathbf{u} \in \mathbf{R}^n$.

Proof: We may assume, without loss of generality, that $\mathbf{v} \cdot (f(\mathbf{x}) - f(\mathbf{z}_0)) \leq 0$ for all $\mathbf{x} \in S(\mathbf{x}_0, r)$. This implies the existence of a point $\mathbf{c} \in B(\mathbf{x}_0, r)$ such that $\mathbf{v} \cdot f(\mathbf{c}) = M$, where $M = \max\{\mathbf{v} \cdot f(\mathbf{x}) : \mathbf{x} \in D(\mathbf{x}_0, r)\}$. Consequently, $\mathbf{v} \cdot f'(\mathbf{c})\mathbf{u} = 0$ for all $\mathbf{u} \in \mathbf{R}^n$. QED

Remark 3. In the case when $n = p = 1$ Theorem 2 says that if for some $z \in (a, b)$ we have

$$\text{j) either } f(z) \geq \max\{f(a), f(b)\} \quad \text{jj) or } f(z) \leq \min\{f(a), f(b)\},$$

then there exists $c \in (a, b)$ such that $f'(c) = 0$. Notice that every $z \in (a, b)$ satisfies either j) or jj) when $f(a) = f(b)$.

The following result (see Boas [3]) is an easy consequence of the above remark.

Corollary 4. Let $a < b$ and $f: [a, b] \rightarrow \mathbf{R}$ be continuous on $[a, b]$ and differentiable on $[a, b]$. Assume that $f'(a) = f'(b)$. Then there exists a point $c \in (a, b)$ such that

$$f'(c)(c - a) = f(c) - f(a).$$

Proof: A straightforward computation shows that Corollary 4 is true for f if and only if it is true for $g(x) = f(x) - xf'(a)$. Therefore we may assume, without loss of generality, that $f'(a) = f'(b) = 0$. Define

$$h(x) = \begin{cases} \frac{f(x) - f(a)}{x - a} & x \neq a \\ 0 & x = a \end{cases}.$$

The function h is continuous on $[a, b]$, differentiable on $(a, b]$ and $h'(b) = -h(b)/(b - a)$.

Assume that $h(b) \neq 0$. From $h(b)h'(b) < 0$ and $h(a) = 0$ we derive the existence of $z \in (a, b)$ which satisfies either i) or ii). In the case when $h(b) = 0$ ($= h(a)$) every point $z \in (a, b)$ will do the job. Hence, by Theorem 2 (Remark 3), there exists $c \in (a, b)$ such that $h'(c) = 0$, and this implies the stated result. QED

Geometrical Applications of Theorem 1 and Theorem 2. We present three geometrical applications. To allow for a visual representation of the results we do not state them in their full generality.

Application 1. Let $f: D(\mathbf{0}, 1) \subset \mathbf{R}^2 \rightarrow \mathbf{R}^3$, $f(u, v) = (x(u, v), y(u, v), z(u, v))$ be continuous on $D(\mathbf{0}, 1)$ and differentiable on $B(\mathbf{0}, 1)$ and let $G = \text{Im}f$. Assume that there exists a plane $p: ax + by + cz + d = 0$, such that $(x(u, v), y(u, v), z(u, v)) \in p$ for every $(u, v) \in S(\mathbf{0}, 1)$. Then there is a point $(u_0, v_0) \in B(\mathbf{0}, 1)$ such that the tangent plane to the surface G at the point $f(u_0, v_0)$ is parallel to p .

Justification. By Theorem 1 (see Remark 1) the vector $\mathbf{v}_0 = (a, b, c)$ is orthogonal to

$$\frac{\partial f}{\partial u}(\mathbf{u}_0) = \mathbf{p} \quad \text{and} \quad \frac{\partial f}{\partial v}(H\mathbf{u}_0) = \mathbf{q},$$

for some $\mathbf{u}_0 \in B(\mathbf{0}, 1)$, $\mathbf{u}_0 = (u_0, v_0)$. The tangent plane to G at $f(\mathbf{u}_0)$ is $\{f(\mathbf{u}_0) + m\mathbf{p} + n\mathbf{q} : m, n \in \mathbf{R}\}$, which is obviously parallel to p .

Application 2. Let $f: D(\mathbf{0}, 1) \subset \mathbf{R}^2 \rightarrow \mathbf{R}^3$, $f(u, v) = (x(u, v), y(u, v), z(u, v))$ be continuous on $D(\mathbf{0}, 1)$ and differentiable on $B(\mathbf{0}, 1)$. Denote by G the surface

$G = \text{Im}f$ and let $G_0 = f(\partial D(0, 1))$. Assume that there is a plane $p: ax + by + cz + d = 0$, such that G_0 is on one side of p and there is a point of S on the other side of p . Then the tangent plane to G at some point $P \in S$ is parallel to p .

Justification. Let $\mathbf{u}_i = (u_i, v_i) \in B(\mathbf{0}, 1)$ be such that $f(\mathbf{u}_i)$ is on the other side of p with respect to G_0 . Then $(a, b, c) \cdot (f(\mathbf{u}) - f(\mathbf{u}_i))$ does not change sign on $\partial D(\mathbf{0}, 1)$. The conclusion follows from Theorem 2.

We illustrate this situation with an example

Example 2. Let $f(u, v) = (u^2 + v^2 - u, u^2 + v, u^2 - v)$. Then $G_0 = \{(1 - u, u^2 + v, u^2 - v) : u^2 + v^2 = 1\}$ and $f(0, 0) = (0, 0, 0)$ are on opposite sides of the plane $p: x + y + z = 1/2$. Hence there is a point P on $G = \text{Im}f$ where the tangent plane is parallel to p . The point is $P = f(1/6, 0)$.

Application 3. Let $\mathbf{x}: [a, b] \rightarrow \mathbf{R}^3$ be continuous on $[a, b]$ and differentiable on $[a, b]$, and let $P = \mathbf{x}(a) = (x(a), y(a), z(a))$, $Q = \mathbf{x}(b) = (x(b), y(b), z(b))$. Then for every plane p passing through the line L joining P with Q there is a point $c \in (a, b)$ such that the vector $\mathbf{x}'(c)$ is parallel to p . In particular, when the plane p is the one containing the origin, we obtain that $\mathbf{x}'(c)$ satisfies the equality

$$\begin{aligned} \text{i)} \quad & x(a)[y(b)z'(c) - z(b)y'(c)] + y(a)[z(b)x'(c) - x(b)z'(c)] \\ & + z(a)[x(b)y'(c) - y(b)x'(c)] = 0. \end{aligned}$$

Justification. The first part is an immediate consequence of Theorem 1, since for every plane passing through L there is a vector \mathbf{u} orthogonal to L and to the plane. For the second part observe that the direction \mathbf{v} of a line orthogonal to p is given by the cross product of the two vectors $\mathbf{x}(a)$ and $\mathbf{x}(b)$, i.e. $\mathbf{v} = \mathbf{x}(a) \times \mathbf{x}(b)$. Thus there exists $c \in (a, b)$ such that $\mathbf{x}(a) \times \mathbf{x}(b) \cdot \mathbf{x}'(c) = 0$, which implies i). For a different justification of the result presented in Application 3 see [2].

Open problem and conjecture. We conclude the paper with an open problem and a conjecture. Theorem 1 and Theorem 2 remain valid if \mathbf{R}^p is replaced by a Hilbert space \mathbf{H} . No changes are needed in the proof. They are also true when \mathbf{R}^p is replaced by a Banach space \mathbf{F} with the vector \mathbf{v} substituted by a linear continuous functional ϕ .

We conjecture that the theorems are false when \mathbf{R}^n is replaced by an infinite-dimensional Banach space \mathbf{E} , because Proposition 2, which plays a key role in both proofs, fails in \mathbf{E} . In fact, the unit closed ball $D(\mathbf{0}, 1)$ of \mathbf{E} is not compact. Consequently, there exists continuous functions $f: D(\mathbf{0}, 1) \rightarrow \mathbf{R}$ such that $\text{Im}f$ is an open interval, as illustrated by the following example.

Example 3. Let \mathbf{H} be the Hilbert space of square summable sequences of real numbers and let D be the disk of \mathbf{H} centered at the origin and with radius 1, $D = D(\mathbf{0}, 1)$. Define

$$T: D \rightarrow \mathbf{H}, T(\mathbf{x}) = T(x_1, x_2, \dots) = (\sqrt{1 - \|\mathbf{x}\|^2}, x_1, x_2, \dots).$$

The map T does not have any fixed point on D . In fact, since $\|T(\mathbf{x})\| = 1$ for all $\mathbf{x} \in D$, every potential fixed point \mathbf{x} must be located on the boundary of D , i.e. \mathbf{x} is fixed for T only if $\|\mathbf{x}\| = 1$. This implies $T(\mathbf{x}) = (0, x_1, \dots)$. Combining this result

with the equality $T(\mathbf{x}) = \mathbf{x}$ gives $\mathbf{x} = \mathbf{0}$, against the assumption $\|\mathbf{x}\| = 1$. The fixed-point free map T allows us to define the continuous function

$$f: D \rightarrow \mathbf{R}, f(\mathbf{x}) = \frac{1}{\|\mathbf{x} - T(\mathbf{x})\|}.$$

Let us show that the image of f is the open half-line $(0.5, \infty)$.

We already know that $\|\mathbf{x} - T(\mathbf{x})\| > 0$ for every $\mathbf{x} \in D$. To verify that the greatest lower bound (glb) of $\{\|\mathbf{x} - T(\mathbf{x})\|: \mathbf{x} \in D\}$ is 0 consider the elements $\mathbf{x}_n \in D(\mathbf{0}, 1)$ whose entries after the n position are all 0, while the first n are all equal to $1/\sqrt{n}$:

$$\mathbf{x}_n = \left(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}}, 0, \dots \right).$$

Clearly $\|\mathbf{x}_n\| = 1$ and $\|\mathbf{x}_n - T(\mathbf{x}_n)\| = \sqrt{2/n}$. Hence the greatest lower bound is 0.

To see that $\|\mathbf{x} - T(\mathbf{x})\| < 2$ for every $\mathbf{x} \in D$, notice that $\|\mathbf{x} - T(\mathbf{x})\| = 2$ requires $\|\mathbf{x}\| = 1$ and $\mathbf{x} = -T(\mathbf{x})$, i.e.

$$(x_1, x_2, \dots) = (0, -x_1, -x_2, \dots).$$

The above equality implies $\mathbf{x} = \mathbf{0}$, a contradiction with $\|\mathbf{x}\| = 1$. To verify that the least upper bound (lub) of $\{\|\mathbf{x} - T(\mathbf{x})\|: \mathbf{x} \in D\}$ is 2 consider the elements \mathbf{y}_n whose entries after the n position are all 0, while the first n are alternatively equal to $\pm 1/\sqrt{n}$:

$$\mathbf{y}_n = \left(\frac{1}{\sqrt{n}}, -\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, -\frac{1}{\sqrt{n}}, \dots, (-1)^{n+1} \frac{1}{\sqrt{n}}, 0, \dots \right).$$

Then $\|\mathbf{y}_n - T(\mathbf{y}_n)\| = \sqrt{4 - (2/n)}$, which implies that the least upper bound is 2. Hence the image of f is the open half-line $(0.5, \infty)$.

It would be nice to have an example which shows that Theorems 1 and 2 fail in infinite dimension. So far we have been unable to construct it.

In the References we mention other contributions (see [5] pg. 19, [6], [8], [9], [11]) regarding Rolle's Theorem, the Mean Value Theorem and the Cauchy Generalized Mean Value Theorem. They are not directly related to this paper, but the reader may find them useful to get a better overview of the work done in this area.

ACKNOWLEDGMENT. We would like to thank the referee for many valuable comments, which have greatly improved the paper, and in particular for calling our attention to the result of Sanderson.

REFERENCES

1. Ahlfors L. V. (1953): *Complex Analysis*. McGraw-Hill Book Company, pg. 39.
2. Barret L. C., Jacobson R. A. (1960): Extended Laws of the Mean. *Am. Math. Monthly*, **67**, 1005–1007.
3. Boas, R. B. JR. (1961): *A Primer of Real Functions*. The Carus Mathematical Monographs **13**, MAA, J. Wiley and Sons.
4. Evard J. Cl., Jafari F. (1992): A complex Rolle's Theorem. *Am. Math. Monthly*, **99**, 858–861.
5. Flett T. M. (1980): *Differential Analysis*. Cambridge University Press.
6. Furi M., Martelli M. (1991): On the Mean Value Theorem, Inequality and Inclusion. *Am. Math. Monthly*, **98**, 840–847.
7. Marden M. (1985): The search for a Rolle's Theorem in the complex domains. *Am. Math. Monthly*, **92**, 643–650.
8. McLeod R. M. (1964): Mean Value Theorem for Vector Valued Functions. *Proc. Edinburgh Math. Soc.* **14**, 197–209.

9. Rosenholtz I. (1991): A topological Mean Value Theorem for the plane. *Am. Math. Monthly*, **98**, 149–153.
10. Sanderson D. E. (1972). A versatile Vector Mean Value Theorem. *Am. Math. Monthly*, **79**, 381–383.
11. Tineo A. (1989): A generalization of Rolle's Theorem and an application to a nonlinear equation. *J. Austr. Math. Soc. Ser. A*, **46**, 395–401.

*Dipartimento di Matematica Applicata
Università di Firenze
Via S. Marta 3, Firenze, Italy
furi@ifiudg.bitnet*

*Mathematics Department
CSU Fullerton
Fullerton, CA 92634
mmartelli@hmcvax.claremont.edu*

Reply to CD's

"These are indeed exciting times in the world of Mathematics." I would like to respond to the "Tale of Two CD's" by Dan Kennedy. "The winds of change are blowing through ... the curriculum" and some of us feel like the French citizens in the late 1930's that we might be better off without some of the coming changes. I am a practicing mathematician of a dozen years experience writing simulations, optimizations, and analyses in wireless and landline telephony, printed circuit board production, airline fleet assignment, yield management, and maintenance delays. I also have considerable exposure to Mathematics education as consumer and producer.

It is apt that he chooses the compact (CD) as his analog (pun intended) for the newest New Math. The CD is truly a triumph of marketing over technology. It is quiet and cute, shiny and high-tech. If the medium were truly digital, then the sound wouldn't be dramatically altered by putting a rubber mat on top, painting the rim green, or reversing the prongs of the AC cord. By stuffing thousands of dollars of digital signal processors (DSPs) into the signal path, clever engineers have surpassed cheap turntables to the point where the best \$10,000 CD players outperform \$1000 turntables. But, of course, you're listening to the DSPs rather than the CD.

Those of us who keep concert seats year after year in spite of the surface noise (audience rustling) and clicks and pops (coughs and sneezes) tend also to find ourselves labelled as "collectors" and "Luddites" as we continue to purchase records. I have over two thousand phonograph records and a Linn, LOCI, and EK-1 to play them. The huge advantage of CD over record is the low manufacturing cost which should have brought the consumer cheap recordings, but somehow this never happened.

The educational analogy to "compact" sound is a simplified curriculum relying on technology to replace the drudgery of traditional teaching methods. We are offering better high school mathematics programs than before, alas, to college students and, occasionally, to graduate students. Reducing student involvement in math courses has failed to attract better or more motivated students to our classrooms; did we really expect it to do so?

In our *Brave New World* (Aldous Huxley, 1932) of post-Modern education, the emphasis is on maintaining the students' willingness to enroll in our courses and come to our classes. We must entertain them and we mustn't scare them away so machines do their "timeses and gazintas" and solve equations for them and invert matrices for them and even graph functions for them. Being able to balance a checkbook without a machine is *A Sense of Power* (Asimov, 1957) in today's Mathematics classroom.

Mathematics is not a spectator sport; we learn it by doing it. While my Linear Programming students this fall will learn to use AMPL modeling language, they also will graph polytopes and crank out Simplex optimizations by hand.

Do I suppose Newton would be flattered to see our students walking a road to discovery essentially the same as his? *I certainly do*. I know I'm flattered to see my own discovery process (including my software) used ten years later to teach new students in cellular mobile telephone system engineering as its success in pedagogy affirms my own confidence in my knowledge. Isaac Newton's results are certainly more than thirty times as worthy of posterity as mine and Calculus and Physics students should see them his way.

Our Mathematics education and curriculum certainly could use a dose of enthusiasm and support from both teacher and students, but I doubt it requires much revision. Computational and display tools can enhance and deepen our insights and our delight, but we must remember that students learn by traveling the road to discovery with their own eyes, ears, and limbs and not by watching machines (or professors) do it for them.

Adam N. Rosenberg
14061 Oakgreen Circle South
Afton, MN 55001
Adam@Psionic.mn.org

NOTES

Edited by: John Duncan

Adding Distinct Congruence Classes Modulo a Prime

Noga Alon, Melvyn B. Nathanson and Imre Ruzsa

1. THE ERDŐS-HEILBRONN CONJECTURE. The Cauchy-Davenport theorem [3, 4] states that if A and B are nonempty sets of congruence classes modulo a prime p , and if $|A| = k$ and $|B| = l$, then the sumset

$$A + B = \{a + b \mid a \in A, b \in B\}$$

contains at least $\min(p, k + l - 1)$ congruence classes. It follows that the sumset $A + A$ contains at least $\min(p, 2k - 1)$ congruence classes. Erdős and Heilbronn conjectured 30 years ago that there are at least $\min(p, 2k - 3)$ congruence classes that can be written as the sum of two *distinct* elements of A . Erdős has frequently mentioned this problem in his lectures and papers (for example, Erdős-Graham [6, p. 95]). Applying results from exterior algebra and the representation theory of the symmetric group, Dias de Silva and Hamidoune [5] recently proved this conjecture. The purpose of this paper is to give a simple proof of the Erdős-Heilbronn conjecture that uses only the most elementary properties of polynomials. The method, in fact, yields generalizations of both the Erdős-Heilbronn conjecture and the Cauchy-Davenport theorem.

2. THE POLYNOMIAL METHOD

Lemma 1 (Alon-Tarsi [2]). *Let A and B be nonempty subsets of a field F with $|A| = k$ and $|B| = l$. Let $f(x, y)$ be a polynomial with coefficients in F and of degree at most $k - 1$ in x and $l - 1$ in y . If $f(a, b) = 0$ for all $a \in A$ and $b \in B$, then $f(x, y)$ is identically zero.*

Proof: This follows immediately from the fact that a nonzero polynomial $p(x) \in F[x]$ of degree at most $k - 1$ cannot have k distinct roots in F . We can write

$$f(x, y) = \sum_{i=0}^{k-1} \sum_{j=0}^{l-1} f_{i,j} x^i y^j = \sum_{i=0}^{k-1} v_i(y) x^i,$$

where

$$v_i(y) = \sum_{j=0}^{l-1} f_{i,j} y^j$$

is a polynomial of degree at most $l - 1$ in y . Fix $b \in B$. Then

$$u(x) = f(x, b) = \sum_{i=0}^{k-1} v_i(b) x^i$$

is a polynomial of degree at most $k - 1$ in x such that $u(a) = 0$ for all $a \in A$. Since $u(x)$ has at least k distinct roots, it follows that $u(x)$ is the zero polynomial, and so $v_i(b) = 0$ for all $b \in B$. Since $\deg(v_i) \leq l - 1$ and $|B| = l$, it follows that $v_i(y)$ is the zero polynomial, and so $f_{i,j} = 0$ for all i and j . This completes the proof. \square

Lemma 2. *Let A be a finite subset of a field F , and let $|A| = k$. For every $m \geq 0$ there exists a polynomial $r_m(x) \in F[x]$ of degree at most $k - 1$ such that*

$$r_m(a) = a^m$$

for all $a \in A$.

Proof: We shall give two proofs. Let $A = \{a_0, a_1, \dots, a_{k-1}\}$. We must show that there exists a polynomial $r_m(x) = z_0 + z_1x + \dots + z_{k-1}x^{k-1} \in F[x]$ such that

$$r_m(a_i) = z_0 + z_1a_i + z_2a_i^2 + \dots + z_{k-1}a_i^{k-1} = a_i^m$$

for $i = 0, 1, \dots, k - 1$. This is a system of k linear equations in the k unknowns z_0, z_1, \dots, z_{k-1} , and it has a solution if the determinant of the coefficients of the unknowns is nonzero. The Lemma follows immediately from the observation that this determinant is the Vandermonde determinant

$$\begin{vmatrix} 1 & a_0 & a_0^2 & \cdots & a_0^{k-1} \\ 1 & a_1 & a_1^2 & \cdots & a_1^{k-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & a_{k-1} & a_{k-1}^2 & \cdots & a_{k-1}^{k-1} \end{vmatrix} = \prod_{0 \leq i < j \leq k-1} (a_j - a_i) \neq 0.$$

The second proof is even simpler. Let

$$t(x) = \prod_{j=0}^{k-1} (x - a_j).$$

Then $t(x) \in F[x]$ and $\deg(t) = k$. By the division algorithm for polynomials over a field, for every $m \geq 0$ there exist polynomials $q_m(x)$ and $r_m(x)$ such that $\deg(r_m) \leq k - 1$ and

$$x^m = t(x)q_m(x) + r_m(x).$$

Then

$$a_i^m = t(a_i)q_m(a_i) + r_m(a_i) = r_m(a_i)$$

for all $a_i \in A$. This completes the proof. \square

Theorem 1. *Let p be a prime number, and let $F = \mathbf{Z}/p\mathbf{Z}$. Let A and B be nonempty subsets of the field F such that $|A| \neq |B|$. Let*

$$C = \{a + b \mid a \in A, b \in B, a \neq b\}.$$

Then

$$|C| \geq \min(p, |A| + |B| - 2).$$

Proof: Let $|A| = k$ and $|B| = l$. We can assume that

$$1 \leq l < k \leq p.$$

If $k + l - 2 > p$, let $l' = p - k + 2$. Then

$$2 \leq l' < l < k$$

and

$$k + l' - 2 = p.$$

Choose $B' \subseteq B$ such that $|B'| = l'$, and let

$$C' = \{a + b' | a \in A, b' \in B', a \neq b'\}.$$

Then $C' \subseteq C$. If the Theorem holds for the sets A , B' , and C' , then

$$|C| \geq |C'| \geq k + l' - 2 = p = \min(p, |A| + |B| - 2).$$

Therefore, we can assume that

$$k + l - 2 \leq p.$$

We must prove that

$$|C| \geq k + l - 2.$$

Suppose that

$$|C| \leq k + l - 3.$$

Choose w so that

$$w + |C| = k + l - 3.$$

We construct the polynomial $f(x, y)$ in $F[x, y]$ as follows: Let

$$f(x, y) = (x - y)(x + y)^w \prod_{c \in C} (x + y - c).$$

Then f has total degree exactly $k + l - 2$, and

$$f(a, b) = 0 \quad \text{for all } a \in A, b \in B.$$

Moreover,

$$\begin{aligned} f(x, y) &= \sum_{\substack{i, j \geq 0 \\ i+j \leq k+l-2}} f_{i,j} x^i y^j \\ &= (x - y)(x + y)^{k+l-3} + \text{lower order terms.} \end{aligned}$$

Since $1 \leq l < k \leq p$ and $1 \leq k + l - 3 \leq p - 1$, it follows that the coefficient $f_{k-1, l-1}$ of the monomial $x^{k-1}y^{l-1}$ in $f(x, y)$ is

$$\binom{k+l-3}{k-2} - \binom{k+l-3}{k-1} = \frac{(k-l)(k+l-3)!}{(k-1)!(l-1)!} \not\equiv 0 \pmod{p}.$$

By Lemma 2, for every $m \geq k$ there exists a polynomial $r_m(x)$ of degree at most $k - 1$ such that $r_m(a) = a^m$ for all $a \in A$, and for every $n \geq l$ there exists a polynomial $s_n(y)$ of degree at most $l - 1$ such that $s_n(b) = b^n$ for all $b \in B$. We use the polynomials $r_m(x)$ and $s_n(y)$ to construct a new polynomial $f^*(x, y)$ from $f(x, y)$ as follows: If $x^m y^n$ is a monomial in $f(x, y)$ with $m \geq k$, then we replace $x^m y^n$ with $r_m(x) y^n$. Since $\deg(f) = k + l - 2$, it follows that if $m \geq k$, then

$n \leq l - 2$, and so $r_m(x)y^n$ is a sum of monomials $x^i y^j$ with $i \leq k - 1$ and $j \leq l - 2$. Similarly, if $x^m y^n$ is a monomial in $f(x, y)$ with $n \geq l$, then we replace $x^m y^n$ with $x^m s_n(y)$. If $n \geq l$, then $m \leq k - 2$, and so $x^m s_n(y)$ is a sum of monomials $x^i y^j$ with $i \leq k - 2$ and $j \leq l - 1$. This determines a new polynomial $f^*(x, y)$ of degree at most $k - 1$ in x and $l - 1$ in y . The process of constructing $f^*(x, y)$ from $f(x, y)$ does not alter the coefficient $f_{k-1, l-1}$ of the term $x^{k-1} y^{l-1}$, since this monomial does not occur in any of the polynomials $r_m(x)y^n$ or $x^m s_n(y)$. On the other hand,

$$f^*(a, b) = f(a, b) = 0$$

for all $a \in A$ and $b \in B$. It follows immediately from Lemma 1 that the polynomial $f^*(x, y)$ is identical zero. This contradicts the fact that the coefficient $f_{k-1, l-1}$ of $x^{k-1} y^{l-1}$ in $f^*(x, y)$ is nonzero, and completes the proof. \square

Theorem 2 (Dias da Silva-Hamidoune). *Let p be a prime number, and let $F = \mathbf{Z}/p\mathbf{Z}$. Let $A \subseteq F$, and let $|A| = k \geq 2$. Let $2^{\wedge}A$ denote the set of all sums of two distinct elements of A . Then*

$$|2^{\wedge}A| \geq \min(p, 2k - 3).$$

Proof: Let $A \subseteq F$, $|A| \geq 2$. Choose $a \in A$, and let $B = A \setminus \{a\}$. Then $|B| = |A| - 1$. Let

$$C = \{a + b | a \in A, b \in B, a \neq b\}.$$

Then $2^{\wedge}A \supseteq C$ and, by Theorem 1,

$$|2^{\wedge}A| \geq |C| \geq \min(p, |A| + |B| - 2) = \min(p, 2|A| - 3).$$

This completes the proof of the Erdős-Heilbronn conjecture. \square

Let $k + l - 2 \leq p$, $1 \leq l < k \leq p$. Let $A = \{0, 1, 2, \dots, k - 1\}$ and $B = \{0, 1, 2, \dots, l - 1\}$. Then

$$C = \{a + b | a \in A, b \in B, a \neq b\} = \{1, 2, \dots, k + l - 2\}$$

and

$$2^{\wedge}A = \{1, 2, \dots, 2k - 3\}.$$

This example shows that the lower bounds in Theorem 1 and Theorem 2 are sharp.

3. FURTHER APPLICATIONS OF THE METHOD. The polynomial method is a powerful new technique to obtain results in additive number theory. For example, it gives the following simple proof of the Cauchy-Davenport theorem.

Theorem 3 (Cauchy-Davenport). *Let p be a prime number, and let $F = \mathbf{Z}/p\mathbf{Z}$. Let A and B be nonempty subsets of the field F , and let*

$$C = A + B = \{a + b | a \in A, b \in B\}.$$

Then

$$|C| \geq \min(p, |A| + |B| - 1).$$

Proof: Let $|A| = k$ and $|B| = l$. We can assume that $k + l - 1 \leq p$. If $|C| \leq k + l - 2$, choose w so that

$$w + |C| = k + l - 2,$$

and consider the polynomial

$$f(x, y) = (x + y)^w \prod_{c \in C} (x + y - c).$$

Then $f(a, b) = 0$ for all $a \in A$ and $b \in B$. The polynomial has total degree $k + l - 2$, and the coefficient of the monomial $x^{k-1}y^{l-1}$ is exactly

$$\binom{k + l - 2}{k - 1} \not\equiv 0 \pmod{p}.$$

The proof proceeds exactly as the proof of Theorem 1. \square

As a final example of the method, we state and prove the following new result.

Theorem 4. *Let A and B be nonempty subsets of $F = \mathbf{Z}/p\mathbf{Z}$, and let*

$$C = \{a + b \mid a \in A, b \in B, ab \neq 1\}.$$

Let $|A| = k$ and $|B| = l$. Then

$$|C| \geq \min(p, k + l - 3).$$

Proof: If $k + l - 3 > p$, let $l' = p - k + 3$. Then $3 \leq l' < l$. Choose $B' \subseteq B$ such that $|B'| = l'$ and let

$$C' = \{a + b' \mid a \in A, b' \in B', ab' \neq 1\}.$$

Since $C' \subseteq C$, it suffices to prove that $|C'| \geq k + l' - 3$. Equivalently, we can assume that $k + l - 3 \leq p$, and we must prove that $|C| \geq k + l - 3$.

Suppose that $|C| \leq k + l - 4$. Choose w so that

$$w + |C| = k + l - 4,$$

and consider the polynomial

$$f(x, y) = (xy - 1)(x + y)^w \prod_{c \in C} (x + y - c).$$

Then $f(a, b) = 0$ for all $a \in A$ and $b \in B$. The polynomial has total degree $k + l - 2$, and the coefficient of the monomial $x^{k-1}y^{l-1}$ is

$$\binom{k + l - 4}{k - 2} \not\equiv 0 \pmod{p}.$$

The proof continues exactly as the proof of Theorem 1. \square

Let $k + l - 1 \leq p$, let $k, l \geq 2$, and choose $d \in \mathbf{Z}/p\mathbf{Z}$, $d \neq 0$, such that

$$(1 + (k - 1)d)(1 + (l - 1)d) = 1.$$

Let $A = \{1, 1 + d, 1 + 2d, \dots, 1 + (k - 1)d\}$ and $B = \{1, 1 + d, 1 + 2d, \dots, 1 + (l - 1)d\}$. Define C as in Theorem 4. Then $C = \{2 + id \mid i = 1, \dots, k + l - 3\}$. This example shows that the lower bound in Theorem 4 is sharp for $k, l \geq 2$. If $k = 1$, the correct lower bound is $|B| - 1 = k + l - 2$.

4. REMARKS. The results in this paper hold for addition of finite subsets of any field F , where p denotes the characteristic of F if the characteristic is a prime number, and $p = \infty$ if the characteristic is zero.

Dias da Silva and Hamidoune [5] proved the following generalization of the Erdős-Heilbronn conjecture for h -fold sums: Let $h \geq 2$, and let $h^{\wedge}A$ denote the set of all sums of h distinct elements of A . If $A \subseteq \mathbb{Z}/p\mathbb{Z}$ and $|A| = k$, then

$$|h^{\wedge}A| \geq \min(p, hk - h^2 + 1).$$

This result can also be proved by the polynomial method, and we shall present this and other results in a subsequent paper [1].

Nathanson [9] contains proofs of the Cauchy-Davenport theorem and some of its generalizations, as well as a full exposition of the original Dias da Silva-Hamidoune proof of the Erdős-Heilbronn conjecture for h -fold sums, and the polynomial proof. Partial results on the Erdős-Heilbronn conjecture had previously been obtained by Rickert [11], Mansfield [8], Rödseth [12], Pyber [10], and Freiman, Low, and Pitman [7].

REFERENCES

1. N. Alon, M. B. Nathanson, and I. Z. Ruzsa. The polynomial method and restricted sums of congruence classes, *J. Number Theory*, to appear.
2. N. Alon and M. Tarsi. Colorings and orientations of graphs. *Combinatorica*, 12:125–134, 1992.
3. A. L. Cauchy. Recherches sur les nombres. *J. École polytech.*, 9:99–116, 1813.
4. H. Davenport. On the addition of residue classes. *J. London Math. Soc.*, 10:30–32, 1935.
5. J. A. Dias da Silva and Y. O. Hamidoune. Cyclic spaces for Grassmann derivatives and additive theory. *Bull. London Math. Soc.*, 26: 1994, page 140–146.
6. P. Erdős and R. L. Graham. *Old and New Problems and Results in Combinatorial Number Theory*. L'Enseignement Mathématique, Geneva, 1980.
7. G. A. Freiman, L. Low, and J. Pitman. The proof of Paul Erdős' conjecture of the addition of different residue classes modulo prime number. In *Structure Theory of Set Addition, 7–11 June 1993, CIRM Marseille*, pages 99–108, 1993.
8. R. Mansfield. How many slopes in a polygon? *Israel J. Math.*, 39:265–272, 1981.
9. M. B. Nathanson. *Additive Number Theory: 2. Inverse Theorems and the Geometry of Sumsets*. Graduate Texts in Mathematics, Springer-Verlag, New York, 1995.
10. L. Pyber. On the Erdős-Heilbronn conjecture. Personal communication.
11. U.-W. Rickert. *Über eine Vermutung in der additiven Zahlentheorie*. Ph.D. thesis, Tech. Univ. Braunschweig, 1976.
12. Ö. J. Rödseth. Sums of distinct residues mod p . *Acta Arith.*, 65:181–184, 1993.

ALON:
Institute for Advanced Study
Princeton, NJ 08540
and
Department of Mathematics
Tel Aviv University
Tel Aviv, Israel
noga@math.tau.ac.il

RUZSA:
Mathematical Institute
of the Hungarian Academy of Sciences
Budapest, P.O.B. 127
H-1364, HUNGARY
h1140ruz@ella.hu

NATHANSON:
Department of Mathematics
Lehman College (CUNY)
Bronx, NY 10468
nathansn@dimacs.rutgers.edu

A Simple Proof of the Hölder and the Minkowski Inequality

Lech Maligranda

The proofs as well as the extensions, inverses and applications of the well-known Hölder and Minkowski inequalities can be found in many books about real functions, analysis, functional analysis or L_p -spaces (cf. [Mi]). The aim of this note is to give another proof of these classical inequalities. The following lemma will be a main step in our simple proof of these inequalities. This lemma was motivated by considerations in [KPS], [M] and [MP].

Lemma. For $1 \leq p < \infty$ and any $a, b > 0$, we have

$$(i) \quad \inf_{t>0} \left[\frac{1}{p} t^{1/p-1} a + \left(1 - \frac{1}{p} \right) t^{1/p} b \right] = a^{1/p} b^{1-1/p}.$$
$$(ii) \quad \inf_{0<t<1} \left[t^{1-p} a^p + (1-t)^{1-p} b^p \right] = (a+b)^p.$$

First proof. In these proofs we will use calculus.

(i) Let, for $t > 0$, the function f be defined by

$$f(t) = \frac{1}{p} t^{1/p-1} a + \left(1 - \frac{1}{p} \right) t^{1/p} b.$$

Then the derivative f' satisfies

$$f'(t) = \frac{1}{p} \left(\frac{1}{p} - 1 \right) t^{1/p-2} a + \left(1 - \frac{1}{p} \right) \frac{1}{p} t^{1/p-1} b = \frac{1}{p} \left(\frac{1}{p} - 1 \right) t^{1/p-2} (a - tb),$$

and so f' is negative for $t < t_0 = a/b$, zero for $t = t_0$ and positive for $t > t_0$. Hence, f has its minimum at the point $t_0 = a/b$ and this minimum is equal to

$$f(t_0) = f\left(\frac{a}{b}\right) = \frac{1}{p} \left(\frac{a}{b}\right)^{1/p-1} a + \left(1 - \frac{1}{p} \right) \left(\frac{a}{b}\right)^{1/p} b = a^{1/p} b^{1-1/p}.$$

(ii) Let, for $0 < t < 1$, the function g be defined by

$$g(t) = t^{1-p} a^p + (1-t)^{1-p} b^p.$$

Then the derivative g' satisfies the equation

$$g'(t) = (1-p)t^{-p}a^p - (1-p)(1-t)^{-p}b^p = 0$$

only when $t = t_1 = a/(a+b)$. Since

$$g''(t) = (1-p)(-p)t_1^{-p-1}a^p - (1-p)(-p)(1-t_1)^{-p-1}b^p > 0,$$

it follows that g has its local minimum at $t_1 = a/(a + b)$, which is equal to

$$\begin{aligned} g(t_1) &= g\left(\frac{a}{a+b}\right) = \left(\frac{a}{a+b}\right)^{1-p} a^p + \left(1 - \frac{a}{a+b}\right)^{1-p} b^p \\ &= \left(\frac{a}{a+b}\right)^{1-p} a^p + \left(\frac{b}{a+b}\right)^{1-p} b^p = (a+b)^p. \end{aligned}$$

This local minimum of the function g is equal to its global minimum because g is continuous on $(0, 1)$ and $\lim_{t \rightarrow 0^+} g(t) = \lim_{t \rightarrow 1^-} g(t) = +\infty$.

Second proof. In these proofs we will use convexity of some functions.

(i) The function $\varphi(u) = \exp(u)$ is convex on R . Thus

$$\begin{aligned} a^{1/p} b^{1-1/p} &= [t^{1/p-1} a]^{1/p} [t^{1/p} b]^{1-1/p} \\ &= \exp \left[\frac{1}{p} \ln(t^{1/p-1} a) + \left(1 - \frac{1}{p}\right) \ln(t^{1/p} b) \right] \\ &\leq \frac{1}{p} \exp[\ln(t^{1/p-1} a)] + \left(1 - \frac{1}{p}\right) \exp[\ln(t^{1/p} b)] \\ &= \frac{1}{p} t^{1/p-1} a + \left(1 - \frac{1}{p}\right) t^{1/p} b \end{aligned}$$

for every $t > 0$. For $t = a/b$ we have equality.

(ii) The function $\psi(u) = u^p$ for $p > 1$ is convex on $[0, \infty)$. Therefore,

$$\begin{aligned} (a+b)^p &= \left[t \frac{a}{t} + (1-t) \frac{b}{1-t} \right]^p \\ &\leq t \left(\frac{a}{t} \right)^p + (1-t) \left(\frac{b}{1-t} \right)^p = t^{1-p} a^p + (1-t)^{1-p} b^p \end{aligned}$$

for every $0 < t < 1$. For $t = a/(a+b)$ we have equality.

Remark 1. If $0 < p < 1$ and we change in the equalities (i) and (ii) the infimum into supremum, then our Lemma is still true.

Remark 2. The second proof of (i) gives also a different proof of the arithmetic-geometric mean inequality

$$a^{1/p} b^{1-1/p} \leq \frac{1}{p} a + \left(1 - \frac{1}{p}\right) b$$

(put $t = 1$) as well as a different proof of the Young inequality

$$ab \leq \frac{1}{p} a^p + \left(1 - \frac{1}{p}\right) b^{1/(1-1/p)}.$$

The classical *Hölder inequality* states: Let $1 \leq p < \infty$ and $1/p + 1/q = 1$. If $x \in L_p(\mu)$ and $y \in L_q(\mu)$, then $xy \in L_1(\mu)$ and

$$(HI) \quad \|xy\|_1 \leq \|x\|_p \|y\|_q.$$

Equivalently, if $x, y \in L_1(\mu)$, then $|x|^{1/p} |y|^{1-1/p} \in L_1(\mu)$ and

$$(HI_1) \quad \left\| |x|^{1/p} |y|^{1-1/p} \right\|_1 \leq \|x\|_1^{1/p} \|y\|_1^{1-1/p}.$$

Proof: According to our Lemma the inequality

$$a^{1/p}b^{1-1/p} \leq \frac{1}{p}t^{1/p-1}a + \left(1 - \frac{1}{p}\right)t^{1/p}b$$

holds for all $t > 0$ and it follows that

$$\begin{aligned} \| |x|^{1/p}|y|^{1-1/p} \|_1 &= \int_{\Omega} |x(s)|^{1/p} |y(s)|^{1-1/p} d\mu(s) \\ &\leq \int_{\Omega} \left[\frac{1}{p}t^{1/p-1}|x(s)| + \left(1 - \frac{1}{p}\right)t^{1/p}|y(s)| \right] d\mu(s) \\ &= \frac{1}{p}t^{1/p-1} \int_{\Omega} |x(s)| d\mu(s) + \left(1 - \frac{1}{p}\right)t^{1/p} \int_{\Omega} |y(s)| d\mu(s) \\ &= \frac{1}{p}t^{1/p-1}\|x\|_1 + \left(1 - \frac{1}{p}\right)t^{1/p}\|y\|_1. \end{aligned}$$

Taking the infimum over all $t > 0$ and using our Lemma again we obtain

$$\| |x|^{1/p}|y|^{1-1/p} \|_1 \leq \|x\|_1^{1/p}\|y\|_1^{1-1/p},$$

which proves inequality (HI₁).

Remark 3. Our proof of (HI₁) still works for a general Banach function space $X(\mu)$ instead of the $L_1(\mu)$ -space, i.e., if $x, y \in X(\mu)$, then $|x|^{1/p}|y|^{1-1/p} \in X(\mu)$ and

$$(HI_X) \quad \| |x|^{1/p}|y|^{1-1/p} \|_X \leq \|x\|_X^{1/p}\|y\|_X^{1-1/p}.$$

Equivalently (cf. [MP]), if $|x|^p \in X(\mu)$ and $|y|^q \in X(\mu)$, $1/p + 1/q = 1$, then $xy \in X(\mu)$ and

$$(HI) \quad \|xy\|_X \leq \| |x|^p \|_X^{1/p} \| |y|^q \|_X^{1/q}.$$

The classical *Minkowski inequality* states: Let $1 \leq p < \infty$. If $x, y \in L_p(\mu)$, then $x + y \in L_p(\mu)$ and

$$(MI) \quad \|x + y\|_p \leq \|x\|_p + \|y\|_p.$$

Proof: By using the second part of our Lemma, i.e. the inequality

$$(a + b)^p \leq t^{1-p}a^p + (1 - t)^{1-p}b^p$$

we find that for all $t, 0 < t < 1$,

$$\begin{aligned} \|x + y\|_p^p &= \int_{\Omega} |x(s) + y(s)|^p d\mu(s) \leq \int_{\Omega} [|x(s)| + |y(s)|]^p d\mu(s) \\ &\leq \int_{\Omega} [t^{1-p}|x(s)|^p + (1 - t)^{1-p}|y(s)|^p] d\mu(s) \\ &= t^{1-p} \int_{\Omega} |x(s)|^p d\mu(s) + (1 - t)^{1-p} \int_{\Omega} |y(s)|^p d\mu(s) \\ &= t^{1-p}\|x\|_p^p + (1 - t)^{1-p}\|y\|_p^p. \end{aligned}$$

Taking the infimum over $0 < t < 1$ and using our Lemma again we obtain

$$\|x + y\|_p^p \leq (\|x\|_p + \|y\|_p)^p,$$

which is inequality (MI).

REFERENCES

[KPS] Krein, S. G., Petunin, Y. U., Semenov, E. M., *Interpolation of Linear Operators*, AMS, Providence 1980.
[M] Maligranda, L., *Calderón-Lozanovskii spaces and interpolation of operators*, Semesterbericht Funktionalanalysis, Tübingen 8 (1985), 83–92.
[MP] Maligranda, L., Persson, L. E., *Generalized duality of some Banach function spaces*, Indagationes Math. 51 (1989), 323–338.
[Mi] Mitrinovic, D. S., *Analytic Inequalities*, Springer-Verlag 1970.

Department of Mathematics
Luleå University
S-971 87 Luleå, Sweden
lech@sm.luth.se

Never Too Late

There is a slip in Williamson’s excellent article [2]. Although nearly 50 years have gone by it is never too late to restore the article to perfection.

The 8 by 8 determinant displayed on page 433 has the value 44, instead of 56 as stated.

From the material in the article that immediately precedes one can deduce that 56 is attainable by appropriately bordering the incidence matrix of the seven point projective plane:

$$\begin{vmatrix} 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \end{vmatrix}$$

Some followup thoughts arise at once but this is not the place to explore them. Years later Ehlich and Zeller [1] showed that 56 is the largest possible value for the determinant of an 8 by 8 matrix consisting entirely of 0’s and 1’s.

1. H. Ehlich and K. Zeller, *Binäre Matrizen*, Zeit Angew. Math. Mech. 42(1962), pages 20–21 of the Sonderheft.
2. J. Williamson, *Determinants whose elements are 0 and 1*, Amer. Math. Monthly 53(1946), 427–434.

Irving Kaplansky
Mathematical Sciences Research Institute
1000 Centennial Drive
Berkeley, CA 94720

Missing Real Numbers

Christopher J. Van Wyk

Authors of programs that implement mathematical algorithms soon confront the fact that computers do not have real arithmetic. Instead, most computers offer some form of floating-point arithmetic, which can be extremely fast, but is also a paltry substitute for real. Floating-point representation can be understood as scientific notation with a limited number of significant figures. Every floating-point system has a smallest $\varepsilon = 2^{-k}$ such that $1 + \varepsilon > 1$; thus, $(1 + \frac{\varepsilon}{2}) + \frac{\varepsilon}{2} = 1$ while $1 + (\frac{\varepsilon}{2} + \frac{\varepsilon}{2}) = 1 + \varepsilon > 1$, so floating-point addition is not even associative. Each floating-point arithmetic operation may commit a relative error of ε ; over a long sequence of operations, such compounded errors can lead to wildly incorrect answers. An especially acute problem is catastrophic cancellation: the computed difference between two floating-point numbers that are nearly equal may include no significant figures at all; this makes it impossible in general to use floating-point arithmetic to test for exact equality [4]. Sometimes people are tempted to ignore all this and to use floating-point arithmetic as if it were real; numerical analysts have long railed against such “naive” use of floating-point [2].

In computing with geometric objects, one promising approach that has emerged is to confine all (would-be) real arithmetic to a few functions and procedures. The rest of the program manipulates discrete information—whether logical, combinatorial, or topological—and calls on the arithmetic subprograms only when necessary. In effect, these subprograms behave as “black boxes”: various complicated and ungainly things may happen inside them, but the rest of the program need not be concerned about the details. Of course, this begs the question of how to write arithmetic subprograms that are robust and reliable.

This column presents some *exact-integer generalized predicates* as examples of arithmetic subprograms. These predicates accept exact integers (usually user data) as input, compute the value of one or more integer polynomials in these inputs, and return as output a member of a small discrete set. For the first example, the set is {true, false}, so it is a predicate even in the ungeneralized sense. The computation of that example uses the generalized predicate sign () that returns one of {+1, 0, -1}, depending on the sign of its argument.

Since the integer arithmetic native to computers also offers a limited number of digits of precision, we shall assume that some multiprecision integer arithmetic is used to compute the values of the polynomials when our integers outgrow the native precision. The many subroutine packages available to perform multiprecision integer arithmetic share at least one property: all run considerably slower than native integer or floating-point arithmetic. The slowdown worsens as the multi-

precision integers get longer; multiplication time, for example, grows quadratically with increasing bitlength. Thus, we shall account carefully for the bitlengths required by our computations.

Our first example of a predicate takes two line segments given by their four endpoints, $(a_i, b_i), (c_i, d_i)$, $i = 0, 1$, and tells whether they intersect transversally (i.e., not at their endpoints). An obvious but naive approach is to compute the rational intersection point $(p/w, q/w)$, where p , q , and w are integers, as the solution to the simultaneous system

$$(c_i - a_i)(y - b_i) = (d_i - b_i)(x - a_i), \quad i = 0, 1;$$

the line segments intersect transversally if and only if $\text{sign}(p - wa_i) = \text{sign}(wc_i - p) \neq 0$ and $\text{sign}(q - wb_i) = \text{sign}(wd_i - q) \neq 0$ for $i = 0, 1$. In general, this computation requires arithmetic in integers whose bitlength is at least three times that of the inputs a_i through d_i .

A better approach is to test whether the endpoints of the first segment lie on opposite sides of the second segment, and vice versa. This amounts to testing whether

$$\text{sign} \begin{vmatrix} a_i & b_i & 1 \\ c_i & d_i & 1 \\ a_{1-i} & b_{1-i} & 1 \end{vmatrix} = -\text{sign} \begin{vmatrix} a_i & b_i & 1 \\ c_i & d_i & 1 \\ c_{1-i} & d_{1-i} & 1 \end{vmatrix} \neq 0 \quad (1)$$

for $i = 0, 1$, which requires arithmetic on integers only about twice as long as the inputs.

The determinant that appears in Eq. 1 of the line-segment intersection test is a special case of the “three-point orientation” predicate, which returns 1 if the points (x_1, y_1) , (x_2, y_2) , and (x_3, y_3) are oriented counterclockwise, -1 if they are oriented clockwise, and 0 if they are collinear. This predicate can be implemented as

$$\text{sign} \begin{vmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \end{vmatrix}. \quad (2)$$

This is the only arithmetic primitive needed to compute the convex hull of a set of points, and it is useful in many algorithms on points and line segments [1].

Let us pause to consider just when Eq. 2 needs to be evaluated exactly. When the magnitude of the determinant is large enough, its sign can be read reliably from a floating-point approximation. When the magnitude is small, however, the floating-point approximation may have the wrong sign, and it is impossible to tell if the determinant is exactly zero. Recognizing the last condition is particularly important for computational geometry, since a zero determinant may well signal some kind of “degeneracy” in the input data; an example of degeneracy for Eq. 2 is three collinear input points.

To enable users to take advantage of some of these observations, thus getting the benefits of exact arithmetic without always having to pay the cost, Steve Fortune (AT & T Bell Laboratories) and I have implemented a program that generates arithmetic subprograms [3]. The user writes exact-integer predicates and specifies the precision of the input integers. Our program compiles each predicate into a $C++$ function that behaves as follows: first, evaluate the expression in floating-point arithmetic; return the sign if the expression is large enough that it is

known to be reliable; otherwise, evaluate the expression exactly in multiprecision integer arithmetic and return the sign.

We have used our program to implement several geometric algorithms. When degeneracies are few (so that the floating-point approximation usually suffices), the resulting programs run almost as fast as programs that use naive floating-point arithmetic. When the input data contains many degeneracies (as it often does in real life), the programs run more slowly than corresponding naive floating-point versions, but they are *correct*: they do not crash ignominiously because conclusions drawn from approximate answers have led to geometrically or topologically impossible situations.

So far, all of our experimental programs have manipulated “flat” objects like points, line segments, lines, and planes. To see how practical and convenient exact-integer arithmetic predicates would be for curved objects, we thought about implementing predicates on circles. This quickly led us to consider a generalization of the three-point orientation predicate to circles: Given four circles $(x - h_i)^2 + (y - k_i)^2 = r_i^2$, $i = 0, 1, 2, 3$, in what order do the six intersection points of circles 1, 2, and 3 with circle 0 appear around that circle?

The obvious approach is to compute the coordinates of the six points directly. As we shall see, the direct algebraic approach is no better an idea here than it was before, so we merely indicate the outlines of the derivation. Begin by computing the line through both of the points at which circle i and circle 0 intersect:

$$a_i = 2(h_i - h_0) \quad (3)$$

$$b_i = 2(k_i - k_0) \quad (4)$$

$$c_i = h_0^2 + k_0^2 + r_1^2 - h_1^2 - k_1^2 - r_0^2 \quad (5)$$

$$a_i x + b_i y + c_i = 0 \quad (6)$$

Next, we can solve Eq. 6 for y in terms of x , substitute into the equation for circle 0, and solve for the x -coordinates of the crossing points; in general, these will be quadratic irrationalities of the form $(A + \sqrt{B})/C$. To express an arithmetic comparison between two such quadratic irrationalities as an integer polynomial, we need to rearrange extensively and square both sides twice (all the while accounting for signs). Evaluating the final integer inequality requires arithmetic in integers at least 20 times as long as the original h_i , k_i , and r_i .

Another way to compute the three-circle orientation predicate is to use some ideas from plane geometry [5]. Define the following:

1. A_i , the *radical axis* of circles 0 and i , is the line given by Eq. 6. This line is perpendicular to the line through (h_0, k_0) and (h_i, k_i) . Orient A_i so that (h_0, k_0) lies to its right.
2. X_i^- and X_i^+ , the two points at which circle i crosses circle 0, lie on A_i . Label the points so that X_i^- precedes X_i^+ along the oriented A_i .
3. C_{ij} , the *radical center* of circles 0, i , and j , is the intersection of A_i and A_j . We note that C_{ij} also lies on the radical axis of circles i and j .
4. P_i is the perpendicular projection of (h_0, k_0) onto A_i .

The radical axes A_i form an arrangement of oriented lines in two-dimensional space, which we can compute explicitly. To compute the three-circle orientation, however, we need to know how the six crossing points X_i^- and X_i^+ lie on the three radical axes. The following observations allow us to determine the positions

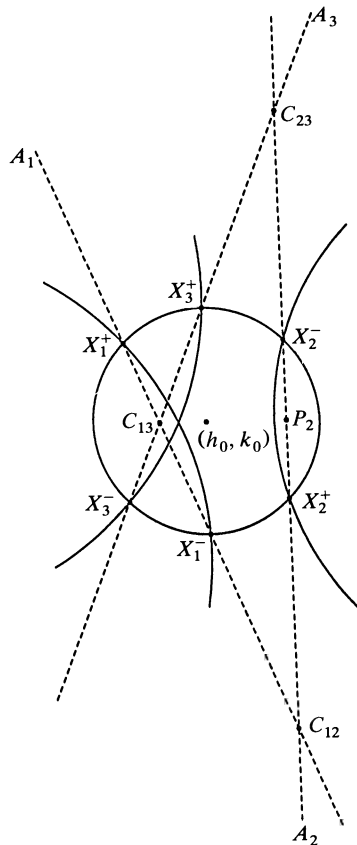


Figure 1. An example of observations (1) and (2) about radical centers and the orientation of crossings. The complete circle is circle 0; only arcs of the other circles are shown.

of the crossings relative to the radical centers without computing their coordinates explicitly:

1. If C_{ij} lies inside circle 0, then C_{ij} lies between X_i^- and X_i^+ on A_i . (For an example, take $i = 1$ and $j = 3$ in Figure 1.)
2. If C_{ij} and C_{jk} both lie outside circle 0, and if P_j lies between C_{ij} and C_{jk} on A_j then x_j^- and x_j^+ lie between C_{ij} and C_{jk} on A_j . (For an example, take $i = 1$, $j = 2$, and $k = 3$ in Figure 1.)
3. If C_{ij} and C_{jk} both lie outside circle 0, and if P_j lies to one side of both C_{ij} and C_{jk} , then X_j^- and X_j^+ both lie to that same side of C_{ij} and C_{jk} on A_j .

Applying these observations allows us to determine which of a finite number of possible arrangements of crossing points X_i^\pm and oriented lines A_i we have. In other words, even if we erased all of the arcs in Figure 1, we could still determine the parts of the dashed lines on which each of the X_i^\pm lie, from which we could deduce the order in which the six X_i^\pm appear around circle 0, all without computing any of the coordinates of X_i^\pm .

To accomplish this feat, we need only construct six topological arcs such that the following three properties hold: (1) Each arc begins at one crossing point and ends at another. (2) The six arcs together form a single, simple cycle. (3) No arc crosses any of the lines A_i . If the boundary of a region in the arrangement

contains only two crossing points, like many of the infinite regions in Figure 1, the choice of arc is forced by requirements (1) and (3). If a region has more than two crossing points on its boundary, like the triangle in Figure 1, the choice of arcs is determined by requirements (1) and (2). In either case, the critical observation is that the choice of arc depends only on topological relationships, not on the actual coordinates of X_i^\pm .

To implement this version of the three-circle predicate, we need to compute quantities like

1. $C_{ij} = (p_{ij}/w_{ij}, q_{ij}/w_{ij})$, where

$$p_{ij} = \begin{vmatrix} -c_i & b_i \\ -c_j & b_j \end{vmatrix}, \quad q_{ij} = \begin{vmatrix} a_i & -c_i \\ a_j & -c_j \end{vmatrix}, \quad w_{ij} = \begin{vmatrix} a_i & b_i \\ a_j & b_j \end{vmatrix}, \quad (7)$$

assuming the definitions of a_i , b_i , and c_i given by Eqs. 3–5;

2. whether the radical center C_{ij} is inside circle 0:

$$\text{sign}\left((p_{ij} - h_0 w_{ij})^2 + (q_{ij} - k_0 w_{ij})^2 - r^2 w_{ij}^2\right); \quad (8)$$

3. the order of C_{ij} , C_{jk} , and P_j along A_j ; we test this by testing the signs of C_{ij} and C_{jk} in the following expression, whose zero-locus is the line perpendicular to A_j through the center of circle 0:

$$(y - k_0)(p_{ij} w_{jk} - p_{jk} w_{ij}) - (x - h_0)(q_{ij} w_{jk} - q_{jk} w_{ij}). \quad (9)$$

The quantities in Eq. 7 are two to three times the bitlengths of the inputs h_i , k_i , and r_i , while Eqs. 8 and 9 are about six times the bitlengths of the inputs, the latter after some more algebraic simplification.

While this second version of the computation is conceptually more involved, it realizes a substantial reduction in the required bitlength for multiprecision arithmetic operations (from 20 to 6). The savings is considerably greater if the coefficients in the original circle equations are allowed to be rational instead of being restricted to integers. The extensive and intricate geometric reasoning required does, however, call into question the practicality of basing reliable arithmetic exclusively on integer computation.

Another possible foundation for exact computation would be a system for representing algebraic numbers. Here, a number x is represented as a polynomial p and an open interval (a, b) such that $x \in (a, b)$, $p(x) = 0$, and $p(y) \neq 0$ for $y \in (a, b)$, $y \neq x$ [6]. In such a system, we could use the rational parameterization of the circle to compute the parameters of the intersection points as the solutions to quartic equations; sorting these parameters would tell the three-circle orientation. Efforts to make it practical to compute using polynomials and root-isolating intervals are underway [7].

With a sufficiently powerful symbolic-algebra system, one could even write the three-circle orientation test by solving for the coordinates of the intersection points, then sorting some arctangents. Such a simple real formulation of the solution is undeniably appealing, but it is likely to be even more expensive than using algebraic numbers. This dilemma accounts for the title of this column. Not only are the real numbers missing from computer arithmetic: sometimes we find ourselves really missing them.

REFERENCES

1. H. Edelsbrunner, *Algorithms in Computational Geometry*, Berlin: Springer-Verlag, 1987.
2. G. E. Forsythe, Pitfalls in computation, or why a math book isn't enough, *Amer. Math. Monthly* 77:9 (1970), 931–956.
3. S. Fortune and C. J. Van Wyk, Efficient exact arithmetic for computational geometry, *Proc. Ninth Ann. Symp. Comput. Geom.*, 163–172, 1993.
4. D. Goldberg, What every computer scientist should know about floating-point arithmetic, *Comput. Surveys* 21:1 (1991), 5–48.
5. D. Pedoe, *Geometry: A Comprehensive Course*, Cambridge: Cambridge Univ. Press, 1970.
6. J. T. Schwartz and M. Sharir, On the “piano movers” problem. II. general techniques for computing topological properties of real algebraic manifolds, *Advances in Appl. Math.* 4 (1983), 298–351.
7. C. Yap and T. Dube, The exact computation paradigm. In D. Z. Du and F. K. Hwang, eds., *Computing in Euclidean Geometry*, 2d ed., World Scientific, to appear.

Department of Mathematics & Computer Science
Drew University
Madison, NJ 07940
cvanwyk@drew.edu

Applied Mathematics

On page 920 of the November issue of the Monthly, you quote George Steiner saying that “applied mathematics is a higher plumbing, a kind of music for the police band.” Poincaré, who observed that “nature not only poses for us problems, she suggests the method of solution,” would have disagreed violently; so would have Newton, Euler, Lagrange, Riemann, Hilbert, Hadamard, Wiener, Weyl, Birkhoff, von Neumann, Kac and Friedrichs. The only mathematician that comes to mind who expressed such views is G.H. Hardy; “A Mathematician’s Apology” is marred by just such exaltation of the uselessness of mathematics. The great chemist Soddy put him down harshly but justly: “From such cloistral clowning the world sickens.”

November must be snobbery month, for on page 901 you quote Alfred Adler: “Each generation has its few great mathematicians, and mathematics would not even notice the absence of the others.” Evidently Adler is not familiar with the principle of the integral calculus. Let’s hope that in the meanwhile Adler has managed to overcome his inferiority complex.

Peter D. Lax
Courant Insitute of Mathematical Sciences
New York University
251 Mercer Street
New York, NY 10012-1185

THE EVOLUTION OF . . .

Edited by Abe Shenitzer

Mathematics, York University, North York, Ontario M3J 1P3, Canada

The Evolution of Algebra 1800–1870*

I. G. Bashmakova and A. N. Rudakov

The first event of this period was the appearance, in 1801, of C. F. Gauss' *Disquisitiones Arithmeticae*. Of the seven parts of the book only one is devoted to an algebraic issue, namely the cyclotomic equation $x^n - 1 = 0$. But the author's brilliant algebraic thinking is apparent in all the other parts as well. *Disquisitiones*, an epoch-making work in algebraic number theory, was for a long time a handbook and source of ideas in algebra. In the course of his study of the cyclotomic equation Gauss shows that it is solvable for every n in the sense that the solutions are expressible in terms of radicals, gives a method for explicitly finding these expressions, and singles out the values of n for which the solutions are expressible in quadratic radicals and thus the values of n for which it is possible to construct a regular n -gon by means of ruler and compass. As always, his investigations are strikingly profound and detailed. They were continued by N. H. Abel, who proved the insolubility by radicals of the general quintic and singled out a class of equations, now named for him, that are solvable by radicals. The new notions of field (domain of rationality) and group (group of an equation) turned up in Abel's papers with greater definiteness. The next step in this direction that completed the theory was the papers of the young E. Galois, published in fragmentary form between 1830 and 1832, and, after his death, in more complete form by Liouville in 1846.

The papers of Abel, and especially of Galois, already belong to the radically new trend of ideas now generally accepted in algebra. In his study of the ancient problem of solution of equations by radicals Galois shifted the center of gravity from the problem to the methods of its solution: he gave clear-cut definitions of the concepts of a field and of the group of an equation, established the correspondence between the subgroups of the group of an equation and the subfields of the splitting field of the polynomial on the left side of that equation, and, finally, singled out the normal subgroups of a group and studied its composition series. These were completely new and extremely fruitful methods of investigation and yet

*This article is a reprint of the major part of the introduction to an essay dealing with the evolution of algebra and algebraic number theory during the period of 1800–1870. The essay forms Chapter 2 of the book *Mathematics of the 19th Century* that deals with mathematical logic, algebra, number theory, and probability theory in the 19th century. Chapter 2 was written by I. G. Bashmakova and A. N. Rudakov with the assistance of A. N. Parshin and E. I. Slavutin. The book was published in 1992 by Birkhäuser Verlag and is a translation of a Russian book published by Nauka in 1978. (The reprinted material is found on pp. 36–40 of the Birkhäuser Verlag book.) Reprinted with permission.

they were apprehended by mathematicians only in the 70s. The one exception was groups of substitutions. Such groups were considered by Galois and their investigation began already in the 40s.

Another source of group theory was Gauss' theory of composition of classes of forms. In this theory one applied an operation analogous to addition (or multiplication) of numbers to objects very different from numbers. Gauss' study of forms of the same discriminant was in effect a study of the fundamental properties of cyclic and general abelian groups.

The two parts of Gauss' remarkable paper "The theory of biquadratic residues" appeared in 1828 and 1832, respectively. In it Gauss not only gave a geometric interpretation of the complex numbers (this was done before him) but also—and this is very important—transferred to complex numbers the notion of a whole number, a concept that seemed inseparable from the rational integers for more than 2000 years.

Gauss constructed an arithmetic of complex integers entirely analogous to the usual arithmetic and used the new numbers to formulate the law of biquadratic reciprocity. This opened for arithmetic boundless new horizons. Soon Eisenstein and Jacobi formulated and proved the law of cubic reciprocity and used for this purpose numbers of the form $K + m\rho$, $\rho^3 = 1$, $\rho \neq 1$, and in 1846 P. Lejeune-Dirichlet found all units (that is invertible elements) of the ring of integers of the field $\mathbb{Q}(\theta)$, where θ is a root of

$$x^n + a_1x^{n-1} + \cdots + a_n = 0,$$

$a_i \in \mathbb{Z}$.¹ This paper, with its deep results in the theory of algebraic numbers, is also of interest from the point of view of group theory: in it Dirichlet constructed the first nontrivial example of an infinite abelian group and investigated its structure.

Further progress in algebraic number theory was linked to reciprocity laws and to Fermat's last theorem. Attempts to prove this theorem brought E. Kummer to the study of the arithmetic of fields $\mathbb{Q}(\zeta)$, $\zeta^p = 1$, $\zeta \neq 1$. In 1844–1847 Kummer discovered that if one defines a "prime" number to be an indecomposable integer in a field $\mathbb{Q}(\zeta)$, then the law of unique factorization into prime factors fails for the integers in $\mathbb{Q}(\zeta)$. To "save the day" and restore the possibility of constructing an arithmetic analogous to the usual (arithmetic) he introduced ideal factors. In so doing, he laid the foundations for the subtlest and most abstract theories of algebraic number theory. Kummer's methods were local. They were further developed by E. I. Zolotarev, K. Hensel, and others, and now form the core of commutative algebra.

Linear algebra continued to develop in the first half of the 19th century. In this connection, the first thing to be noted is that whereas no part of Gauss' *Disquisitiones* deals directly with linear algebra, its advance was bound to be furthered by the detailed study of integral quadratic forms in two variables contained in that work. A. Cauchy's "On an equation for the determination of the secular inequalities of planetary motions" (1826) dealt implicitly with the eigenvalues of matrices of arbitrary order. Somewhat later, in 1834, there appeared C. G. J. Jacobi's "On the transformation of two arbitrary homogeneous functions of the second order by means of linear substitutions into two others containing only squares of the variables; together with many theorems on the transformation of multiple integrals" in which he explicitly studied quadratic forms and their reduction to

¹ \mathbb{Z} is the ring of integers and \mathbb{Q} is the field of rational numbers.

canonical form. Jacobi also perfected the theory of determinants (1841). What was still lacking in this theory was geometric features and, above all, the all-important and fundamental notion of a linear space. The first, none-too-clear definition of a linear space was given by H. Grassmann in his *Die lineale Ausdehnungslehre* of 1844. This work, rich in new ideas but written in a muddled manner, first attracted attention when its author published a reworked and improved version in 1862. In particular, the work contains a construction of exterior products and the now famous Grassmann algebra. In 1843 there appeared A. Cayley's *Chapters in the analytical geometry of (n) dimensions*, a work less rich in ideas but better known to contemporary mathematicians. There is a close connection between the development of linear algebra and the theory of hypercomplex numbers (now known as the theory of algebras) which elicited considerable interest at the time. Years of fruitless attempts to generalize the complex numbers were crowned with success in 1843 by W. R. Hamilton's discovery of the quaternions. Hamilton studied the quaternions for over 20 years, for the rest of his life. His researches are summarized in two fundamental works: *Lectures on quaternions* (1853) and *Elements of the theory of quaternions* (1866). Their subsequent significance is due not so much to quaternions but to the new notions and methods of "vector calculus" introduced in this connection.

To resume our account of the further development of group theory we mention the series of A. Cauchy's papers, published between 1844–1846, in which he proves a great variety of theorems on groups of substitutions (subgroups of the symmetric group), including the famous theorem of Cauchy to the effect that a group whose order is divisible by a prime p contains an element of order p . A further major event in the history of group theory was the publication—in three parts (1854, 1854, 1859)—of Cayley's paper *On the theory of groups, as depending on the symbolic equation $\theta^n = 1$* . Following the spirit of the English school, Cayley views a group as an abstract set of symbols with a given law of composition and defines a number of fundamental notions of abstract group theory, chief among them being the notions of a group and of isomorphism. This was a notable step in the evolution of the new abstract mathematical thinking.

Of crucial importance for the further development of group theory was the appearance, in 1870, of C. Jordan's fundamental *Traité des substitutions et des équations algébriques*. This work contained the first systematic and complete exposition of Galois theory as well as a detailed presentation of results in group theory up to that time, including Jordan's own significant results in these areas. In it Jordan also introduced what is now known as the Jordan canonical form of matrices of linear transformations. The publication of Jordan's work was a major event in all of mathematics.

Mention must be made of the flourishing, in the middle of the 19th century, of an area of algebra intermediate between linear algebra and algebraic geometry known as the theory of invariants. On the one hand, its content consists in the generalization and development of topics in linear algebra such as reduction to canonical form of quadratic forms and matrices of linear transformations. On the other hand, it is the study, in concrete situations, of the answer to the following question: "Given a geometric object determined in some coordinate system by certain algebraic conditions, find a way of obtaining from the algebraic conditions geometric characteristics of the object that are invariant with respect to coordinate transformations." Between 1840–1870 many of the works of various mathematicians dealt with the determination of systems of invariants in different concrete situations. The best known are the works of Cayley, Eisenstein, Sylvester, Salmon

and Clebsch. In this connection one must single out two papers by Hesse, published in 1844 and 1851, respectively, in which he introduced the notion of a hessian and applied it to geometry, and P. Gordan's famous 1868 paper in which he proved a general algebraic theorem on the existence of a finite system of base invariants. An important paper close to these investigations is Cayley's *A sixth memoir upon quantics* (1859). In it Cayley showed how to consider the metric properties of geometric figures from the single viewpoint of the theory of invariants. This paper was one of the sources of F. Klein's Erlangen Program that resulted in revolutionary changes in geometry.

At that time, an important achievement in linear algebra was Sylvester's 1852 proof of the law of inertia of quadratic forms, presented in the paper *Proof of the theorem that every homogeneous quadratic polynomial can be reduced by means of a real orthogonal substitution to the form of a sum of positive and negative squares*. It was proved, but not published, somewhat earlier by Jacobi. In 1858 there appeared Cayley's *Memoir on the theory of matrices*. In it Cayley introduced the algebra of square matrices and established the isomorphism between the algebra of quaternions and a certain algebra of second-order matrices (a subalgebra of the algebra of all square second-order complex matrices). This work was of great importance for the clarification of the relation between the theory of algebras and linear algebra.

In the sixties, the activities of K. Weierstrass had an important influence on the development of mathematics. He published virtually nothing but included the results of his investigations in his lectures at Berlin University. In his 1861 lectures Weierstrass introduced the notion of a direct sum of algebras and showed that every (finite dimensional) commutative algebra (over the field of real numbers) without nilpotent elements is the direct sum of copies of the fields of real and complex numbers. This was one of the earliest classification results in algebra.

One of the main problems of algebraic number theory in the sixties and seventies was the extension of Kummer's divisibility theory from cyclotomic fields to general algebraic number fields. This was accomplished in three different constructions due, respectively, to E. I. Zolotarev, R. Dedekind, and L. Kronecker. Of the three, it was Dedekind's work—the Xth Supplement to Dirichlet's lectures on number theory published in 1871 and the XIth Supplement to subsequent editions—that was accepted by all mathematicians as the solution of the problem. Dedekind's clear, algebraically transparent, account became the model of mathematical style for many decades to come. By this and other works Dedekind laid the foundations of the contemporary axiomatic presentation of mathematical theories.

In our survey of the evolution of algebra we have not touched on the theory of elliptic and abelian functions—one of the central lines of development of 19th-century mathematics, an area in which Gauss, Abel, Jacobi, Clebsch, Gordan, Weierstrass and many others invested great efforts. In the 19th century this area belonged primarily to analysis, more specifically to the theory of functions of a complex variable, and it was only gradually, especially at the end of the 19th century, that the role of algebraic ideas in it became very significant.

The algebraization of the area began with Dedekind's transfer of his theory, in a joint work with H. Weber (1882), to the field of algebraic functions. This established the deep parallelism between the theories of algebraic numbers and algebraic functions and was the decisive step for an abstract definition of the concepts of field, module, ring, and ideal. From the end of the last century ideas began to flow in the opposite direction, from the theory of algebraic functions to number theory. This resulted in the introduction of p -adic numbers and topology by means

of p -adic metrics. But this is already part of the mathematics of the present century.

The evolution of the ideas, methods, and theories just described resulted in the creation of abstract “modern algebra” and, later, of algebraic geometry whose flourishing we witness today.

Mathematics

Monthly readers may be interested in the following paragraphs from *Smilla's Sense of Snow* by Peter Høeg. (This book was published in the United States in 1993 by Farrar, Straus, and Giroux, a translation from Danish, translated by Tiina, originally published in Copenhagen in 1992.) The narrator, Smilla, is a young woman who loves mathematics and refers many times to mathematics and mathematicians—including Cantor, Dedekind, Fermat, Newton, and Euclid's *Elements*. On pages 112–113, we find the following:

It seems necessary to explain my claustrophobia to him.

“Do you know what the foundation of mathematics is?” I ask. “The foundation of mathematics is numbers. If anyone asked me what makes me truly happy, I would say: numbers. Snow and ice and numbers. And do you know why?”

He splits the claws with a nutcracker and pulls out the meat with curved tweezers.

“Because the number system is like human life. First you have the natural numbers. The ones that are whole and positive. The numbers of a small child. But human consciousness expands. The child discovers a sense of longing, and do you know what the mathematical expression is for longing?”

He adds cream and several drops of orange juice to the soup.

“The negative numbers. The formation of the feeling that you are missing something. And human consciousness expands and grows even more, and the child discovers the in between spaces. Between stones, between pieces of moss on the stones, between people. And between numbers. And do you know what that leads to? It leads to fractions. Whole numbers plus fractions produce rational numbers. And human consciousness doesn't stop there. It wants to go beyond reason. It adds an operation as absurd as the extraction of roots. And produces irrational numbers.”

He warms French bread in the oven and fills the pepper mill.

“It's a form of madness. Because the irrational numbers are infinite. They can't be written down. They force human consciousness out beyond the limits. And by adding irrational number to rational numbers, you get real numbers.”

I've stepped into the middle of the room to have more space. It's rare that you have a chance to explain yourself to a fellow human being. Usually you have to fight for the floor. And this is important to me.

“It doesn't stop. It never stops. Because now, on the spot, we expand the real numbers with imaginary square roots of negative numbers. These are numbers we can't picture, numbers that normal human consciousness cannot comprehend. And when we add the imaginary numbers to the real numbers, we have the complex number system. The first number system in which it's possible to explain satisfactorily the crystal formation of ice. It's like a vast, open landscape. The horizons. You head toward them and they keep receding. That is Greenland, and that's what I can't be without. That's why I don't want to be locked up.”

Because Smilla Jaspersen knows about snow and ice, she is suspicious about the cause of death of her young friend, Isaiah. *Smilla's Sense of Snow* is the story of her investigation and its impact on Smilla, herself. A well-written book; I enjoyed it a lot.

JoAnne S. Growney
Department of Mathematics and Computer Science
Bloomsburg University
Bloomsburg, PA 17815

PROBLEMS AND SOLUTIONS

Edited by:

Richard T. Bumby, Fred Kochman and Douglas B. West

Proposed problems should be sent to the MONTHLY PROBLEMS address given on the inside front cover. Please include solutions, relevant references, etc. Three copies are requested.

Solutions of published problems should arrive before August 31, 1995 at the MONTHLY PROBLEMS address given on the inside front cover. Solutions should be typed with double spacing, including the problem number and the solver's name and mailing address. Two copies suffice. A self-addressed postcard or label should be included if an acknowledgement is desired.

An asterisk () after the number of a problem, or part of a problem, indicates that no solution is currently available. Partial solutions will be useful in such cases. Otherwise, the published solution is likely to be based on a solution which is complete and correct. Of course, an elegant partial solution or a method leading to a more general result is always useful and welcome. In addition, references to other appearances of MONTHLY problems or to solutions of these problems in the literature are also solicited.*

PROBLEMS

10438. *Proposed by Hunter S. Snevily, University of Idaho, Moscow, ID.*

Let $S = n_1, n_2, \dots, n_k$ be a sequence of positive integers with sum n . Suppose that $n < 2k$. Show that, for all q with $1 \leq q \leq n$, there is a subsequence of S with sum q .

10439. *Proposed by Charles Vanden Eynden, Illinois State University, Normal, IL.*

The rational number $1/9$ is an example of a number c in $[0, 1]$ such that the decimal representation of neither c nor \sqrt{c} contains the digit 0. Find an irrational number with the same property.

10440. *Proposed by Marius Cavachi, Constanța, Romania.*

Show that the Euclidean plane cannot be covered with circular disks having mutually disjoint interiors.

10441. *Proposed by Emre Alkan (student), Bosphorus University, İstanbul, Turkey.*

Given $k + 1$ positive real numbers x_0, \dots, x_k and a positive integer n , show that

$$\sum_{\sigma} (x_{\sigma_1} + \dots + x_{\sigma_k})^{-n} \leq k^{-n} \sum_{i=0}^k x_i^{-n},$$

where the sum on the left is taken over the $k + 1$ distinct k -element subsets of $\{x_0, \dots, x_k\}$.

10442. *Proposed by Roger Bielawski, McMaster University, Hamilton, Ontario, Canada.*

Let f be a continuous function from the unit disc D in \mathbb{R}^2 to itself such that:

$f \circ f$ is the identity of D ; and

f is the identity on the unit circle ∂D .

Show that f is the identity on D .

10443. *Proposed by Ernesto Bruno Cossi, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil.*

Let \mathbf{X} be a topological space. Suppose that there is a mapping f from the underlying set of \mathbf{X} to \mathbb{R} such that $\lim_{x \rightarrow v} f(x) = +\infty$ (in the usual topology on \mathbb{R}) for all limit points v of \mathbf{X} . Prove that every compact subset of \mathbf{X} is denumerable.

10444. *Proposed by Richard L. Bishop and Harold G. Diamond, University of Illinois, Urbana, IL.*

Let Ψ denote the class of measurable functions ψ on $[0, 1]$ with $0 \leq \psi(x) \leq x$ for all $x \in [0, 1]$. Let

$$F(\psi) = \int_0^1 \left(\psi(x) - \int_0^1 \psi(t) dt \right)^2 dx$$

and

$$M = \sup \{F(\psi) : \psi \in \Psi\}.$$

Find M and show that the extreme value is attained.

SOLUTIONS

Counting Pairs of Permutations by Binomial Coefficients

10233 [1992, 571]. *Proposed by M.A. Khan, RDSO, Lucknow, India.*

For any odd positive integer $n = 2r - 1$, prove that

$$\sum_{k=0}^n \frac{(-1)^k}{k+1} \binom{n}{k} \sum_{j=0}^{r-1} (-1)^j \binom{n}{j} (r-j)^{n-k} = \frac{n!}{2}.$$

Solution by Richard Holzsager, The American University, Washington, DC. When we reverse the order of summation and substitute $\frac{1}{n+1} \binom{n+1}{k+1}$ for $\frac{1}{k+1} \binom{n}{k}$, the new inner sum of terms

involving k is $\sum_{k=0}^n (-1)^k \binom{n+1}{k+1} (r-j)^{n-k}$. Summing this by the binomial formula converts the original expression to

$$\frac{1}{n+1} \sum_{j=0}^{r-1} (-1)^j \binom{n}{j} ((r-j)^{n+1} - (r-j-1)^{n+1}).$$

We write this as two sums, change j to $j-1$ in the second sum and recombine, using the fact that $\binom{n}{j-1} (r-j)^{n+1} = 0$ when $j=0$ and $j=r$, to obtain

$$\frac{1}{n+1} \sum_{j=0}^{r-1} (-1)^j \left(\binom{n}{j} + \binom{n}{j-1} \right) (r-j)^{n+1} = \frac{1}{n+1} \sum_{j=0}^{r-1} (-1)^j \binom{n+1}{j} (r-j)^{n+1}.$$

The summand here is unchanged when j is replaced by $n+1-j = 2r-j$, and it is 0 when $j=r$. Thus the original sum equals $\frac{1}{2(n+1)} \sum_{j=0}^{n+1} (-1)^j \binom{n+1}{j} (r-j)^{n+1}$. The summation in this expression is the $(n+1)$ st (backward) difference of the polynomial r^{n+1} . Since the $(n+1)$ st difference of any polynomial in r of degree $n+1$ is $(n+1)!$ times its leading coefficient, the original summation equals $\frac{(n+1)!}{2(n+1)} = n!/2$.

Editorial comment. John Henry Steelman proved in a similar manner the more general result that for any nonnegative integers n and r with $r \leq n+1$,

$$\sum_{k=0}^n \frac{(-1)^k}{k+1} \binom{n}{k} \left(\sum_{j=0}^{r-1} (-1)^j \binom{n}{j} (r-j)^{n-k} + \sum_{j=0}^{n-r} (-1)^j \binom{n}{j} (n+1-r-j)^{n-k} \right) = n!.$$

This reduces to the proposed identity when $n = 2r - 1$.

Solved also by J. Anglesio (France), J. C. Binz (Switzerland), P. Bracken (Canada), R. J. Chapman (U. K.), W. Y. C. Chen, J. Fukuta (Japan), W. T. Gan (student, U. K.), H. van Haeringen (The Netherlands), M. E. H. Ismail, N. Komanda, O. P. Lossers (The Netherlands), M. Mócsy (Hungary), K. Perera (student), C. R. Pranesachar (India), V. S. Ryko (Russia), E. Schmeichel, H.-J. Seiffert (Germany), J. H. Steelman, D. Zeilberger (as O. Khayyam & L. Euler), USA Mathematical Olympiad Program, and the University of Wyoming Problem Circle.

Simultaneous Squares

10238 [1992, 674]. *Proposed by David M. Bloom, Brooklyn College of CUNY, Brooklyn, NY.*

(a) Show that there exist infinitely many positive integers a such that both $a+1$ and $3a+1$ are perfect squares.

(b) Let $a_1 < a_2 < \dots$ be the sequence of all solutions of (a). Show that $a_n a_{n+1} + 1$ is also a perfect square.

Solution by Roman Drnovšek (student), Institute of Mathematics, Physics, and Mechanics, Ljubljana, Slovenia. If $a+1 = x^2$ and $3a+1 = y^2$, then $y^2 - 3x^2 = -2$. The hypothesis modulo 4 implies that x and y must be odd, allowing the equation to be rewritten as $(\frac{3x-y}{2})^2 - 3(\frac{y-x}{2})^2 = 1$. Consider only the solutions in which x and y are positive integers. Then, clearly, $x \leq y \leq 3x$, so $u = (3x-y)/2$ and $v = (y-x)/2$ satisfying the Pell equation $u^2 - 3v^2 = 1$ are also positive integers. The smallest positive solution is $(u_1, v_1) = (2, 1)$, and it is well known that the solutions $\{(u_n, v_n)\}$ satisfy $u_n \pm v_n \sqrt{3} = (u_1 \pm v_1 \sqrt{3})^n = (2 \pm \sqrt{3})^n$ for $n \geq 1$. With $\alpha = 2 + \sqrt{3}$ and $\beta = 2 - \sqrt{3}$, we have $u_n = (\alpha^n + \beta^n)/2$ and $v_n = (\alpha^n - \beta^n)/(2\sqrt{3})$. We obtain infinitely many solutions to (a) by letting $a_n = x_n^2 - 1 = (u_n + v_n)^2 - 1 = (\alpha^{2n+1} - 4 + \beta^{2n+1})/6$. For (b), we compute $a_n a_{n+1} + 1 = [(\alpha^{2n+2} - 8 + \beta^{2n+2})/6]^2$. Since the x_n 's are integers, the a_n 's and $a_n a_{n+1} + 1$ are also integers.

Editorial comment. One could also use recurrence relations for the quantities in the problem. For example, $a_{n+2} = 14a_{n+1} - a_n + 8$, so the sequence begins 8, 120, 1680, 23408, Several readers observed that $a_n a_{n+1} + 1 = [(a_n + a_{n+1})/4 - 1]^2$. Since part (a) follows from the reduction of the given condition to a Pell equation, several solvers replaced 3 by an arbitrary non-square m in that part. The proposer notes that the problem arose from a paper submitted to the New York City High School Math Fair (March, 1991), by Ms. Marianna Mayslich of James Madison High School, concerning sets S of positive integers such that, for some fixed integer t , $ab + t$ is a square whenever a, b are distinct elements of S . This problem shows that $\{1, 3, a_n, a_{n+1}\}$ has that property for $t = 1$. Thus, there are infinitely many four-element sets with this property. The challenge is to find a larger set. A related question is whether $a_m a_n + 1$ can be a square when $|m - n| > 1$.

Solved by 80 other readers and the proposer, with one incorrect solution and two incomplete solutions submitted.

A Voter's Paradox with Majority Rule

10252 [1992, 782]. *Proposed by James S. Weber, The University of Illinois, Chicago, IL.*

An election is to be held with V voters who will rank A alternatives. It is said that alternative X is an " M -majority preference" over alternative Y if there are at least M voters who prefer X to Y . A "voter's paradox cycle" is an ordering of the alternatives $a_0, a_1, \dots, a_{A-1}, a_A = a_0$ so that a_i is preferred over a_{i+1} for $0 \leq i \leq A$. Prove that a voter's paradox cycle can exist for M -majority preference if and only if $AM \leq V(A - 1)$.

Solution by Robert High, New York, NY. First suppose such a cycle does exist and $AM > V(A - 1)$. For each a_i , the set S_i of voters with a_i not preferred to a_{i+1} contains strictly fewer than $V - M$ members. Since $M > V(A - 1)/A$, S_i contains strictly less than V/A members. The union of all the S_i thus contains fewer than V members. This says that the intersection of the complements of all the S_i must be non-empty. But this intersection consists of those voters who prefer a_i to a_{i+1} for all i , a contradiction.

Next, assume that $AM \leq V(A - 1)$. Since M is an integer, we must have $M \leq \lfloor V(A - 1)/A \rfloor$. Note that $V = \lceil V/A \rceil + \lfloor V(A - 1)/A \rfloor$. We will construct a set of voter preferences (total orders of A) yielding a cycle. Let a_0, a_1, \dots, a_{A-1} be some ordering of the alternatives, and let $a_A = a_0$. Also fix an ordering of voters. Let $P = \lceil V/A \rceil$. For the first P voters, specify that $a_0 < a_1$; then specify for the next P voters that $a_1 < a_2$; and continue in this way until we run out of voters. Otherwise, let $a_i > a_{i+1}$ for each voter.

Since $P \geq V/A$, we will exhaust the voters before we run out of alternatives. This means that, for each voter, there will be one a_j that is not preferred to a_{j+1} while a_i is preferred to a_{i+1} in all other cases. Thus, the order defined for each voter is consistent. Indeed, it is a total order in which each alternative is preferred to the next, starting from a_{j+1} . But, for each a_i , we have at least $\lfloor V(A - 1)/A \rfloor \geq M$ voters preferring a_i to a_{i+1} . This completes the construction of the voter paradox cycle.

Editorial comment. This result first appeared in the literature of social choice theory in J. Greenberg, "Consistent majority rules over compact sets of alternatives", *Econometrica* 47 (1979), 627–636. The proposer's solution can be found in J. S. Weber, "An elementary proof of the conditions for a generalized Condorcet paradox", *Public Choice* 77 (1993), 415–419.

On January 8, 1993, Robert High died in a rafting accident in Chile. This solution, his last contribution to this Problem Section, was received on January 2, 1993. His other interests included the game of Go, and he had just become President of the American Go Association at that time. A memorial article appears in the Winter 1993 issue of the *Journal* of that organization. He was a devoted follower of this Problem Section; the solution printed above is a reminder of how much he will be missed.

Solved also by D. Callan, R. J. Chapman (U. K.), K. S. Kedlaya (student), O. P. Lossers (The Netherlands), R. Martin (student), R. Powers, K. Rebman, K. B. Reid, E. Schmeichel, Western Maryland College Problems group, University of Wyoming Problem Circle, and the proposer. One incomplete solution was received.

Complex Roots of Special Quartic Polynomials

10253 [1992, 782]. *Proposed by W. Weston Meyer, General Motors Research Laboratories, Warren, MI.*

Show that the quartic equation

$$z^4 - 2cz^3 + 2\bar{c}z - 1 = 0,$$

where c is a complex number with complex conjugate \bar{c} , has a root not on the unit circle $\{z : |z| = 1\}$ if and only if $(\Re c)^{1/3} + (\Im c)^{1/3}i$ lies outside this circle.

Solution I by W. O. Egerland and C. E. Hansen, ARL, Aberdeen Proving Ground, MD. We show that the roots of the given quartic equation lie on the unit circle if and only if c lies inside or on the astroid $x + iy = \cos^3 t + i \sin^3 t$, $0 \leq t \leq 2\pi$. A result of A. Cohn, specialized to the polynomial $p(z) = z^4 - 2cz^3 + 2\bar{c}z - 1$, states that the zeros of $p(z)$ lie on the unit circle and are simple if and only if the zeros of $p'(z) = 2(2z^3 - 3cz^2 + \bar{c})$ lie in $|z| < 1$ or, equivalently, if the zeros of the polynomial $q(z) = \bar{c}z^3 - 3cz + 2$ lie in $|z| > 1$ (see M. Marden, *Geometry of Polynomials*, American Mathematical Society, 1985, p. 206, Exercise 3). An application of Theorem 6.8b, on p. 493 of P. Henrici, *Applied and Computational Complex Analysis, Vol. 1*, Wiley, 1974, shows that this occurs if and only if $c = c_1 + ic_2$ satisfies $|c| < 1$ and $(c_1^2 + c_2^2 - 1)^3 + 27c_1^2c_2^2 < 0$, i.e., if c lies inside the astroid. If c lies on the astroid, then $p(z) = (z - e^{it})^2(z^2 + (ie^{-it} \sin 2t)z - e^{-2it})$. This completes the proof.

Solution II by Anchorage Math Solutions Group, University of Alaska, Anchorage, AK. As in Solution I, let $p(z) = z^4 - 2cz^3 + 2\bar{c}z - 1$. Then, $e^{i\theta}$ is a root of $p(z)$ if and only if

$$e^{-2i\theta} p(e^{i\theta}) = e^{2i\theta} - 2ce^{i\theta} + 2\bar{c}e^{-i\theta} - e^{-2i\theta} = 0. \quad (1)$$

Write $c = x + iy = |c|e^{i\phi}$. Then (1) becomes $2i \sin 2\theta - 4i |c| \sin(\theta + \phi) = 0$, which simplifies to

$$\frac{x}{\cos \theta} + \frac{y}{\sin \theta} = 1. \quad (2)$$

Thus, the roots of $p(z)$ on the unit circle correspond to the values of θ for which the line with intercepts $(\cos \theta, 0)$ and $(0, \sin \theta)$ contains the point (x, y) . These lines are precisely those for which the distance between intercepts is 1.

Without loss of generality, we may assume that (x, y) lies in the first quadrant. An application of the intermediate value theorem gives a value of θ in the second quadrant and one in the fourth quadrant with the required property. Any other solution would be in the first quadrant.

Let D be the minimum distance between intercepts for lines through (x, y) meeting the positive halves of both axes. By calculus, this is found to be $(x^{2/3} + y^{2/3})^{3/2}$. If $D < 1$, the intermediate value theorem gives two admissible values of θ in the first quadrant. We have now found four roots of $p(z)$ on the unit circle. If $D > 1$ the two solutions previously found are easily seen to be the only solutions. Thus $p(z)$ has two roots on the unit circle and two roots off the circle. If $D = 1$, there is one new value of θ . All roots lie on the unit circle in this case, though one of them is a double root. This has also been noted in Solution I.

Solved also by J. Anglesio (France), F. Brulois, R. J. Chapman (U. K.), P. Deiermann, H. S. Gunaratne (Brunei), H. Kappus (Switzerland), K.-W. Lau (Hong Kong), O. P. Lossers (The Netherlands), T. L. McCoy, A. D. Melas (Greece), Y. Nievergelt, N. Passell, D. Tan, Western Maryland College Problems group, University of Wyoming Problem Circle, and the proposer.

More Isogonal Configurations

10293 [1993, 291]. *Proposed by Moshe Rosenfeld, Pacific Lutheran University, Tacoma, WA.*

Suppose four distinct lines through the origin in \mathbb{R}^3 have the property that the six acute angles between pairs of these lines are all equal. Prove that this configuration of four lines is isometric either to the diagonals of a cube or to a configuration of four of the six diagonals of a regular icosahedron.

Solution I by Raphael M. Robinson, University of California, Berkeley, CA. Let the acute angle between pairs of lines be θ . The four lines will intersect the unit sphere in four pairs of opposite points. We may take one point at the north pole. There will then be three other points in the northern hemisphere at latitude $90^\circ - \theta$. The spherical distance between each pair of points will be either θ or $180^\circ - \theta$. Let the differences in longitude of these three points be α_1, α_2 , and α_3 , so that $\alpha_1 + \alpha_2 + \alpha_3 = 360^\circ$. By the law of cosines for a spherical triangle, we have

$$\cos^2 \theta + \sin^2 \theta \cos \alpha_i = \pm \cos \theta.$$

With the plus or minus sign, this leads to

$$\cos \alpha_i = \frac{\cos \theta}{1 + \cos \theta} \quad \text{or} \quad \cos \alpha_i = \frac{-\cos \theta}{1 - \cos \theta}.$$

Two of the values must be equal, say $\cos \alpha_1 = \cos \alpha_2$. Since $\alpha_1 + \alpha_2 < 360^\circ$, this implies that $\alpha_1 = \alpha_2$, hence $\alpha_3 = 360^\circ - 2\alpha_1$.

If α_3 is also equal to α_1 , then each $\alpha_i = 120^\circ$. Only the second equation relating $\cos \alpha_i$ and $\cos \theta$ is satisfied, and we see that $\cos \theta = 1/3$. The lines form the diagonals of a cube.

If α_3 is not equal to α_1 , then we must have

$$\cos \alpha_1 = \frac{\cos \theta}{1 + \cos \theta} \quad \text{and} \quad \cos \alpha_3 = \frac{-\cos \theta}{1 - \cos \theta}$$

or *vice versa*. Since $\cos \alpha_3 = \cos 2\alpha_1$, this leads in the first case to

$$\frac{-\cos \theta}{1 - \cos \theta} = 2 \left(\frac{\cos \theta}{1 + \cos \theta} \right)^2 - 1,$$

which reduces to $5 \cos^2 \theta = 1$, or $\cos \theta = 1/\sqrt{5}$. The second case leads to the the same conclusion. In the first case, we see that $\cos \alpha_1 = (\sqrt{5} - 1)/4$, hence $\alpha_1 = 72^\circ$, so that $\alpha_2 = 72^\circ$ and $\alpha_3 = 216^\circ$. In the second case, we find that $\alpha_3 = 72^\circ$, hence $\alpha_1 = \alpha_2 = 144^\circ$. Both cases produce four of the six diagonals of a regular icosahedron. The two figures are congruent, since any pair of omitted diagonals can be taken into any other pair.

Thus there are exactly two configurations that satisfy the stated conditions.

Solution II by A. N. 't Woord, University of Technology, Eindhoven, The Netherlands. Let v_1, v_2, v_3, v_4 be unit vectors in \mathbb{R}^3 corresponding with the directions of the four lines. Let A be the 3×4 -matrix $(v_1 v_2 v_3 v_4)$. Define $G = A^t A = \left(\langle v_i, v_j \rangle \right)_{i,j=1}^4$. The matrix G has 1 on the diagonal and has $\pm \alpha$ outside the diagonal, where $\alpha = \cos(\phi)$ and ϕ is the angle between any two of the four lines. After a permutation of v_1, \dots, v_4 and multiplying some of the vectors v_1, \dots, v_4 by -1 we may assume that G is one of the following matrices:

$$\begin{pmatrix} 1 & \alpha & \alpha & \alpha \\ \alpha & 1 & \alpha & \alpha \\ \alpha & \alpha & 1 & \alpha \\ \alpha & \alpha & \alpha & 1 \end{pmatrix}, \begin{pmatrix} 1 & \alpha & \alpha & \alpha \\ \alpha & 1 & -\alpha & \alpha \\ \alpha & -\alpha & 1 & \alpha \\ \alpha & \alpha & \alpha & 1 \end{pmatrix}, \begin{pmatrix} 1 & \alpha & \alpha & \alpha \\ \alpha & 1 & -\alpha & -\alpha \\ \alpha & -\alpha & 1 & -\alpha \\ \alpha & -\alpha & -\alpha & 1 \end{pmatrix}.$$

Observe that the rank of G is not greater than the rank of A which is at most 3. It follows that $\det(G) = 0$.

case 1:

$$G = \begin{pmatrix} 1 & \alpha & \alpha & \alpha \\ \alpha & 1 & \alpha & \alpha \\ \alpha & \alpha & 1 & \alpha \\ \alpha & \alpha & \alpha & 1 \end{pmatrix}$$

Now we get $0 = \det(G) = -(\alpha - 1)^3(3\alpha + 1)$, which leads to a contradiction because $0 \leq \alpha < 1$.

case 2:

$$G = \begin{pmatrix} 1 & \alpha & \alpha & \alpha \\ \alpha & 1 & -\alpha & \alpha \\ \alpha & -\alpha & 1 & \alpha \\ \alpha & \alpha & \alpha & 1 \end{pmatrix}$$

Here we get $0 = \det(G) = (\alpha - 1)(\alpha + 1)(5\alpha^2 - 1)$, so $\alpha = 1/\sqrt{5}$ and $\phi \approx 63.4349^\circ$.

case 3:

$$G = \begin{pmatrix} 1 & \alpha & \alpha & \alpha \\ \alpha & 1 & -\alpha & -\alpha \\ \alpha & -\alpha & 1 & -\alpha \\ \alpha & -\alpha & -\alpha & 1 \end{pmatrix}$$

Now we get $0 = \det(G) = -(3\alpha - 1)(\alpha + 1)^3$, so $\alpha = 1/3$ and $\phi \approx 70.5288^\circ$.

The matrix G determines A up to a left-multiplication of an orthogonal matrix. The second case corresponds with four of the six diagonals of an icosahedron. The third case corresponds with the diagonals of a cube.

Editorial comment. With the goal of producing a “proof without words”, Mario Barra submitted some drawings illustrating the addition of a fourth line to a configuration of three lines. The effort is appreciated; however, the words of the selected solutions seemed more convincing.

The title is borrowed from the article, Timothy Murdoch, “Isogonal configurations”, this MONTHLY, April 1993.

Solved also by M. Barra (Italy), V. Božin (student, Yugoslavia), R. J. Chapman (U. K.), I. Kastanas, O. P. Lossers (The Netherlands), A. D. Melas (Greece), H. Morris, A. Pedersen (Denmark), F. Schmidt, R. Stong, M. Vowe (Switzerland), Anchorage Math Solutions Group, and the proposer.

Sums of C -polynomials

10297 [1993, 291]. *Proposed by Zalman Rubinstein, University of Haifa, Haifa, Israel.*

Let $p(x)$ be a polynomial of degree n .

(a) Show that $p(x)$ can be written as a sum of four polynomials $q_0(x), q_1(x), q_2(x), q_3(x)$, each of degree at most n with all roots of all $q_i(x)$ lying on the unit circle $\{x : |x| = 1\}$.

(b)* Is there a polynomial $p(x)$ which can not be expressed as a sum of fewer than 4 such $q_i(x)$?

Solution by Richard Stong, Rice University, Houston, TX. We show that 3 such $q_i(z)$ suffice for any $p(z)$. First note that $f(z) = (2i)^{-n}(z - 1)^n g(i(z + 1)/(z - 1))$ is a polynomial of degree exactly n with all of its roots on $\{z : |z| = 1\}$ if and only if $g(z) = (z - i)^n f((z + i)/(z - i))$ is a polynomial of degree exactly n with all real roots. Therefore it suffices to express any polynomial of degree at most n as a sum of three polynomials of degree exactly n with all real roots.

Let $p(z) = f(z) + ig(z)$ where f and g are polynomials with real coefficients. Choose a constant C with $|f(x)| \leq C$ and $|g(x)| \leq C$ for $x \in [-1, n]$. Let

$$q_0(z) = (1+i)Kz(z-1)\dots(z-n+1)$$

for some large K ($K > 4C$ suffices); let

$$q_1(z) = f(z) - Kz(z-1)\dots(z-n+1)$$

and

$$q_2(z) = i(g(z) - Kz(z-1)\dots(z-n+1)).$$

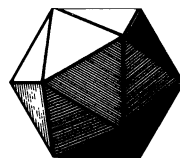
Clearly $p(z) = q_0(z) + q_1(z) + q_2(z)$ and $q_0(z)$ has only real roots. Furthermore, if K is large enough, then the polynomial $Kz(z-1)\dots(z-n+1)$ will exceed C in magnitude at the points $z = -1/2, 1/2, \dots, n-1/2$ and the values at these points will alternate in sign. Since $f(z)$ and $g(z)$ lie in $[-C, C]$ for these z , the polynomials $q_1(z)$ and $q_2(z)$ are not identically zero and each must have a root in each of the intervals $[k-1/2, k+1/2]$, $0 \leq k \leq n-1$. Since they are polynomials of degree at most n , these must be all roots, and the polynomials have exact degree n .

Editorial comment. The solvers listed below all solved part (a); both Richard Stong and Antonios D. Melas showed that *three* polynomials suffice. Thus part (b) has a negative answer, and it is natural to inquire whether two polynomials will suffice. In fact, this question was addressed in the solution of Melas. The idea of the construction is to find a polynomial $G(z)$ of degree 4 that cannot be the sum of two polynomials with real roots, and apply the transformation used in the selected solution with $n = 4$ to obtain $F(z)$. The inverse transformation takes a C -polynomial of degree 4 into a polynomial with only real roots multiplied by a power of $z - i$. This requires a study of $G(z)$ at $z = i$. These considerations lead to choosing $G(z) = z^4 + 4iz + 1$. The fact that $G'(i) = 0$ is used to rule out the possibility of $z = i$ being a multiple root of one of the summands. If $G(z)$ is a sum of two polynomials with only real roots, its zero coefficients will lead to zero coefficients of the terms of degree 2 and 3 in the summands. However, such polynomials cannot have only real roots. The possibility that one of the summands in the representation of $G(z)$ has a simple root at $z = i$ is handled similarly. A detailed proof would be too long to reproduce here.

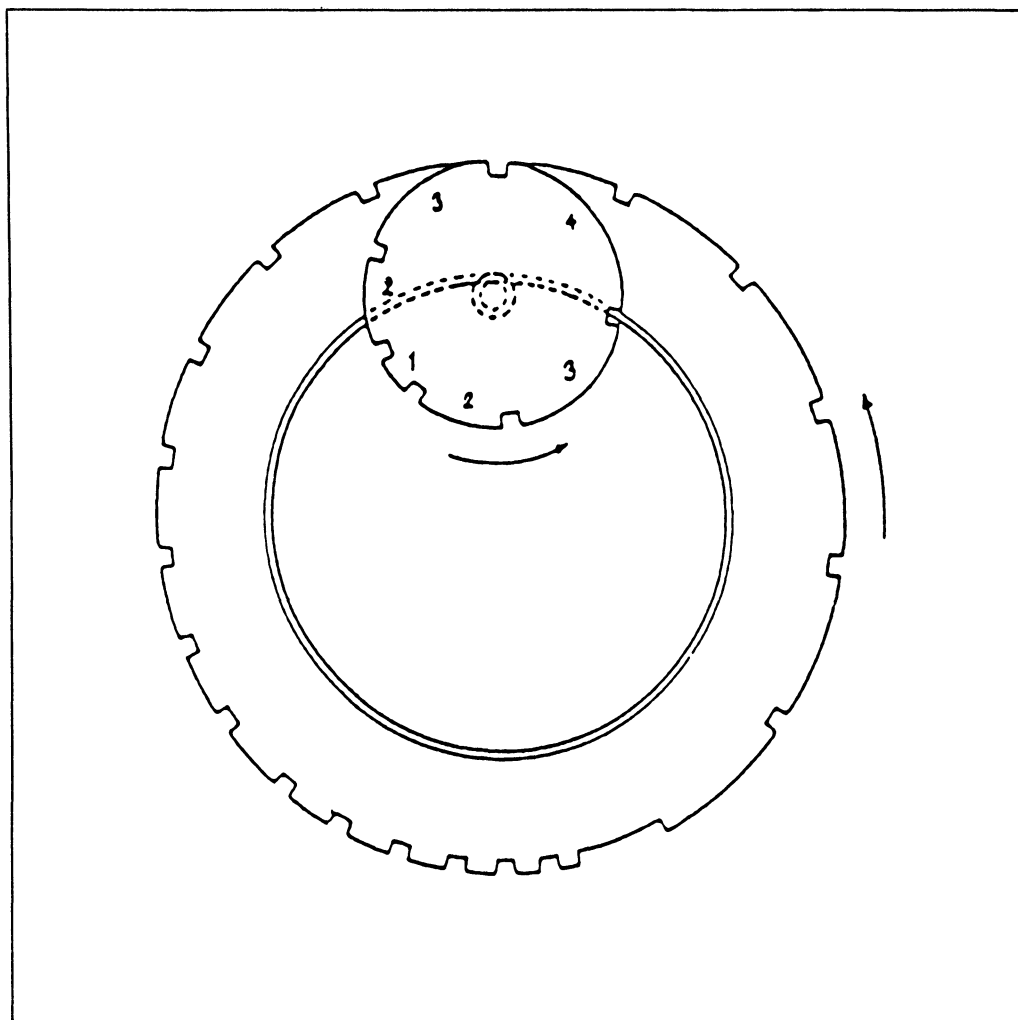
Solved also by I. Kastanas, O. P. Lossers (The Netherlands), F. Schmidt, and the proposer.

Collaborating editors: David F. Appleyard, Paul T. Bateman, Bruce C. Berndt, Duane M. Broline, Barry W. Brunson, Frank S. Cater, Gulbank D. Chakerian, Underwood Dudley, Gerald A. Edgar, Michael A. Filaseta, Ira M. Gessel, Richard A. Gibbs, Jerrold R. Griggs, Douglas A. Hensley, John R. Isbell, Mourad E. H. Ismail, Murray Klamkin, Daniel J. Kleitman, Frederick W. Luttmann, Frank B. Miles, Richard Pfeifer, Stephen L. Portnoy, J. O. Shallit, John Henry Steelman, Kenneth B. Stolarsky, David E. Tepper, Douglas B. Tyler, Daniel Ullman, and William E. Watkins.

The American Mathematical Monthly



Volume 102 Number 4 / APRIL 1995



Count-Wheels
(see page 310)

AN OFFICIAL PUBLICATION OF THE MATHEMATICAL ASSOCIATION OF AMERICA

NOTICE TO AUTHORS

The *Monthly* publishes articles, notes, and other features about mathematics and the profession. The readership of the *Monthly* is intended to include everybody who is mathematically inclined, including of course professional mathematicians and students of mathematics at all collegiate levels. While no single article or feature is likely to appeal to everyone, material should interest and be accessible to a large number of readers. This is the most important criterion for acceptance.

Articles may be expositions of old results or presentations of new ones. They may concern all of mathematics or one small area, a broad development or a single application, historical reminiscences or one important event. While some articles may contain the author's new research, the novelty of material and generality of the results is far less important than the clarity of exposition and general interest. Discussing one illuminating case of a well known result is far better than providing all the details of an obscure but new proposition. Articles in the *Monthly* are supposed to inform and to entertain; they are meant to be read rather than archived.

Notes are short and possibly informal articles. A note may concern a clever new proof of an old theorem, a novel way to present tired material, or a lively discussion of a philosophical (but still mathematical) issue. Also, any topic is suitable, so long as it is related to mathematics. Because a note is short, the first few sentences are the most important part: They should explain the purpose and invite the reader in. Photographs or diagrams often will attract the reader's attention.

All articles and notes should be sent to the editor:

JOHN EWING
Department of Mathematics
Indiana University
Bloomington, IN 47405

Please send 3 copies, typewritten on only one side of the paper. Illustrations should be carefully drawn on separate sheets of paper in black ink; the original should be without lettering and two copies should have appropriate captions and lettering indicated.

Proposed problems or solutions should be sent to:

RICHARD BUMBY,
P.O. Box 10971
New Brunswick, NJ 08906-0971.

Please send 2 copies of all material, typewritten if possible.

Letters to the Editor, both for publication and for private reading, should be sent to the Editor at the address given above. Comments, including criticisms, are welcome, as are all suggestions for making the *Monthly* a lively, entertaining, and informative journal.

EDITOR:

JOHN H. EWING

ASSOCIATE EDITORS:

PETER BORWEIN	FRED KOCHMAN
RICHARD BUMBY	CATHERINE MCGEOCH
DENNIS DETURCK	RICHARD NOWAKOWSKI
UNDERWOOD DUDLEY	ARNOLD OSTEBEE
JOHN DUNCAN	LEE RUBEL
JOAN FERRINI-MUNDY	ABE SHENITZER
JOSEPH GALLIAN	LYNN STEEN
STEVEN GALOVICH	STAN WAGON
RICHARD GUY	DOUGLAS WEST
DARRELL HAILE	HERBERT WILF
PAUL HALMOS	SANDY ZABELL
JOAN HUTCHINSON	PAUL ZORN

EDITORIAL ASSISTANT:

MISTY CUMMINGS

STAFF ARTIST:

MIKE CAGLE

QUOTE MASTER:

MARK WOODARD

Reprint permission:

MARCIA P. SWARD, Executive Director

Advertising Correspondence:

Ms. ELAINE PEDREIRA, Advertising Manager

Subscription correspondence, change of address, and other inquiries:

Membership / Subscriptions Department

All at the address:

The Mathematical Association of America
1529 Eighteenth Street, N.W.
Washington, DC 20036.

Microfilm Editions: University Microfilms International, Serial Bid coordinator, 300 North Zeeb Road, Ann Arbor, MI 48106.

The AMERICAN MATHEMATICAL MONTHLY (ISSN 0002-9890) is published monthly except bimonthly June-July and August-September by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, DC 20036 and Montpelier, VT. Copyrighted by the Mathematical Association of America (Incorporated), 1995, including rights to this journal issue as a whole and, except where otherwise noted, rights to each individual contribution. General permission is granted to Institutional Members of the MAA for noncommercial reproduction in limited quantities of individual articles (in whole or in part) provided a complete reference is made to the source. Second class postage paid at Washington, DC, and additional mailing offices. **Postmaster:** Send address changes to the American Mathematical Monthly, Membership / Subscription Department, MAA, 1529 Eighteenth Street, N.W., Washington, DC, 20036-1385.



Contents

ARTICLES

Gale's Round-Trip Jeep Problem/ALAN HAUSRATH,
BRADLEY JACKSON, JOHN MITCHEM,
and EDWARD SCHMEICHEL 299

Count-Wheels: A Mathematical Problem Arising in Horology/
STEVEN H. WEINTRAUB 310

How to Teach a Class by the *Modified* Moore Method/
DONALD R. CHALICE 317

The Significant-Digit Phenomenon/THEODORE P. HILL 322

Exploring the Brachistochrone Problem/LADAWN HAWS
and TERRY KISER 328

Continued Fractions, Chebychev Polynomials, and Chaos/
WILLIAM DERRICK and JACK EIDSWICK 337

FEATURES

COMMENTS 298

NOTES

A Relation Between Partitions and the Number of Divisors/
WANG ZHENG BING, ROBERT FOKKINK,
and WAN FOKKINK 345

Answers to Two Questions Concerning Quotients of Primes/
PAOLO STARNI 347

Avoiding the Exchange Lemma/JAMES FORD 350

Intervals Contained in Arithmetic Combinations of Sets/
STEPHEN SILVERMAN 351

UNSOLVED PROBLEMS

Does the Möbius Function Determine Multiplicative Arithmetic?/
D. FLATH and A. ZULAUF 354

THE AUTHORS 357

PROBLEMS AND SOLUTIONS 359

REVIEWS

Hilbert's Tenth Problem. By Yuri V. Matiyasevich/
MARTIN DAVIS 366

TELEGRAPHIC REVIEWS 370

Gale's Round-Trip Jeep Problem

Alan Hausrath, Bradley Jackson, John Mitchem,
Edward Schmeichel

In 1947 Fine [Fin] introduced and solved a problem of maximizing the distance a jeep can travel into the desert using n drums of fuel. Subsequently, Phipps [Phi], Alway [Alw], and Gale [Gal] gave other solutions to the original problem or considered related problems. As mentioned in [Fin], the original problem is similar to one which arose in air transport operations in the China theater during World War II, and it has been suggested that there may be applications to Arctic expeditions and interplanetary travel.

Near the end of [Gal], the author states, "An apparently simple question is the round trip problem in which fuel is available at both ends of the desert, but I must confess . . . that I have not been able to find the solution. It is not hard to see that one can do at least as well in this case as in the case of two jeeps making one-way trips, but it may be possible to do better. The difficulty here as with many optimization problems is that there does not appear to be any simple way to determine whether or not a given solution is optimal."

Gale's problem can be interpreted in two equivalent ways. (i) Given unlimited fuel at each end of a desert of given length, find a round trip across the desert which uses as little fuel as possible. (ii) Given a fixed amount of fuel which can be distributed between the two ends of a desert, find the maximum length desert which can be crossed in a round trip using the available fuel. We find it convenient to consider (ii) and give an optimal solution for it. We also describe a solution for the analogous round trip problem where the two allowed depots may be placed anywhere in the desert.

In each of the above problems the jeep can carry exactly 1 drum. It is implicit that the jeep can store whatever fraction of a drum is desired at any point in the desert. (Perhaps the driver carries large plastic bags for fuel storage.) In [Dew], Dewdney proposed an interesting variation of the one-way problem. Although Dewdney's problem was given in terms of drums, gallons, and miles, it can be rephrased as follows: Find the maximum distance a jeep can travel into the desert using n drums of fuel where the jeep can carry 1 drum plus $1/5$ of a drum in its tank, but only drums can be stored. That is the jeep can dump at most $5/6$ of its fuel capacity in the desert. It is interesting to note that Dewdney's problem has been solved as a linear programming problem; an optimal algorithm for Dewdney's problem appears in [Jac]. But the problems solved in this paper apparently are not easily posed as either linear or dynamic programming problems. In [Gal], Gale also points out that "there is a feeling among many people that the original jeep problem can be solved by the functional equation method of dynamic programming . . . I know of no way of solving the problem by this method."

1. A BRIEF DESCRIPTION OF THE SOLUTION TO GALE'S PROBLEM. In solving Gale's problem we will start by considering the longest desert which can be crossed in a round trip if there are m drums of fuel at the start S and k drums of

fuel at the finish F . Let $D(f)$ denote the length of the optimal one-way trip using f drums of fuel. If $m \leq k$, it is clear that one can do no better than $D(m)$ and should use the S -fuel outbound and the F -fuel returning. For $m > k$, going $D(m)$ outbound will not work as the jeep is unable to return to S . Instead, in order to make full use of the drums at S , on the outbound trip a number of depots are created leaving fuel for the return. Let T denote the location of the depot furthest from S . We prove that the following highly plausible qualitative conditions determine an optimal solution: (i) Use only S -fuel when going from S to F . (ii) Use only F -fuel when returning from F to T . (iii) Use only S -fuel stored at the depots, when returning from T to S . The solution then follows by putting together solutions of previously solved jeep problems. Thus it follows from (ii) that the distance from F to T is $D(k)$, and the distance from T to S is obtained by solving a slight variation of the well-known round trip jeep problem with fuel only at S .

To finish Gale's problem, we need only find the optimal distribution of the available fuel between S and F .

2. ORIGINAL PROBLEMS. We have x drums of fuel available at the edge of the desert and a jeep which can carry at most 1 drum. Here we give the well-known algorithm for maximizing the one-way distance, and an algorithm for maximizing the round trip distance for the jeep using x drums of which k must be delivered to F . One unit of distance will be the distance that the jeep can travel on one drum of fuel. We assume that the jeep's efficiency is constant. It does not depend on wind, weather, weight, or depth of the ruts in the sand. The algorithms and their optimality proofs which we give are based on work appearing in [Gal], [Phi], [Fin], and [Niv, Section 10.9].

Theorem A. *Given $n + f$, $0 \leq f < 1$, drums of fuel at the start and a jeep with capacity of 1 drum, the maximum one-way distance which the jeep can travel is*

$$D_1 = 1 + \frac{1}{3} + \frac{1}{5} + \cdots + \frac{1}{2n-1} + \frac{f}{2n+1}.$$

Proof: We begin with an algorithm which achieves distance D_1 . First assume $f = 0$. Repeat n times: Put 1 drum of fuel into the jeep, drive forward $1/(2n-1)$ units, store $1 - (2/(2n-1))$ units, and return to the previous fuel dump, except on the n th iteration do not return. We now have

$$(n-1) \left(1 - \frac{2}{2n-1} \right) + 1 - \frac{1}{2n-1} = n-1$$

drums of fuel, and the jeep at distance $1/(2n-1)$ from the previous dump. Iterate this process, replacing n successively by $n-1, n-2, \dots, 1$.

If $f > 0$, begin the above process by first moving all $n+1$ drums forward $f/(2n+1)$ units, thus delivering n full drums to the first fuel dump.

In order to show that D_1 is the maximum attainable distance when $f = 0$, for any integer i , $0 \leq i \leq n$, we let x_i denote the point on segment SF such that the total distance traveled on the right of x_i is i . (We emphasize that the distance traveled is on the right, not 'toward' the right.) Then $x_n = S$ and $x_0 = F$. Since the jeep used exactly k units of fuel while traveling on the right of x_k , at least k units must arrive at x_k . Let P be any point between x_{k+1} and x_k .

Then more than k units of fuel must have crossed P , and thus the jeep crossed P while traveling toward the right at least $k+1$ times. It follows that the jeep

crossed P at least k times while going toward the left. Hence the jeep crossed P at least $2k + 1$ times and the distance between x_{k+1} and x_k , denoted (x_{k+1}, x_k) is at most $1/(2k + 1)$. Thus we have that the distance from S to F is at most

$$\begin{aligned} & (x_n, x_{n-1}) + (x_{n-1}, x_{n-2}) + \cdots + (x_1, x_0) \\ & \leq \frac{1}{2n-1} + \frac{1}{2n-3} + \cdots + \frac{1}{3} + 1 = D_1. \end{aligned}$$

Similarly when $f > 0$, the distance from S to x_n is at most $f/(2n + 1)$. This completes the proof of Theorem A.

Theorem B. Let m, k be integers, $0 \leq f, g < 1$, and $m + g > k + f$. Given $m + g$ drums of fuel at S and a jeep with capacity 1 drum, the maximum round trip distance in which the jeep delivers $k + f$ drums to F is

$$D_2 = \begin{cases} \frac{g-f}{2m+2}, & \text{if } m = k; \\ \frac{g}{2m+2} + \frac{1}{2m} + \frac{1}{2m-2} + \cdots + \frac{1}{2k+4} + \frac{1-f}{2k+2}, & \text{if } m \geq k+1. \end{cases}$$

Proof: First consider the case when $m \geq k + 1$. For any integer i , $m + g \geq i \geq k + f$, let x_i be the point such that exactly i units of S -fuel are used to the right of that point or delivered to F . Let P be a point between x_i and x_{i-1} , for $m \geq i \geq k + 2$. Then P is crossed at least $2i$ times using S -fuel and $\text{dist}(x_i, x_{i-1}) \leq 1/2i$. Also if P is between S and x_m , then P is crossed at least $2m + 2$ times using S -fuel so that $\text{dist}(S, x_m) \leq g/(2m + 2)$. Similarly if P is between x_{k+1} and F , then P is crossed at least $2k + 2$ times using S -fuel and $\text{dist}(x_{k+1}, F) \leq (1 - f)/(2k + 2)$. Thus the distance between S and F is at most D_2 .

We now give an algorithm which uses $m + g$ drums of fuel for a round trip of length D_2 and delivers $k + f$ drums of that fuel to F . At S , repeat $m + 1$ times: Put $(m + g)/(m + 1)$ units into the jeep, drive forward $g/(2m + 2)$ units and leave all fuel except just enough to return to S . On the last trip do not return to S but leave all fuel at this dump. The total amount of fuel at the dump at this stage is $(m + g) - (2m + 1)(g/(2m + 2)) = m + (g/(2m + 2))$. Leave $g/(2m + 2)$ for the return journey. For each i , $m \geq i \geq k + 2$, fill the jeep i times and each time go forward $1/2i$ units. On each of the first $i - 1$ times dump $1 - (1/i)$ units and return. The last time dump $1 - (1/2i)$ units. Now we use $i - 1$ units to go on to the next fuel dump, having left $1/2i$ for the return. Finally, use $1 - f$ units to make $k + 1$ round trips between x_{k+1} and F . This delivers $(k + 1) - (1 - f) = k + f$ units to F at distance $(1 - f)/(2k + 2)$ from x_{k+1} . This completes the proof when $m \geq k + 1$. The proof when $m = k$ is analogous.

We observe that when $k + f = 0$, Theorem B gives the well-known maximum length desert which can be crossed in a round trip using only fuel from S .

3. ROUND TRIPS WITH 2 FUEL DEPOTS. We consider two different problems, one in which we have the depots at each end of the desert, and one in which we can put the second depot at any point. Of course the first depot must be at the start. In both problems we want to maximize the length of the desert which can be crossed for a fixed amount of fuel.

We begin with a theorem which gives the maximum length desert which can be crossed given that some fixed amounts of fuel are available at each end of the desert.

Suppose that a total of x drums of fuel are available. One way to proceed is to divide the fuel equally between S and F . Then use the one-way algorithm for traveling from S to F and for returning to S from F . We show, however, that it is more efficient to allocate less than half the fuel to F . In fact the maximum distance is achieved when F receives only $k = \lfloor ((x + 1)/2)^{1/2} \rfloor$ drums of fuel and $x - k$ drums are available at S .

If $x - k_1(k_1)$ drums of fuel are available at $S(F)$, where $k_1 < x/2$, Theorem 1 establishes the maximum desert length that can be crossed. The algorithm for attaining that distance establishes a point T between S and F . The distance from T to F is the maximum one-way distance which can be crossed using k_1 drums of fuel. The jeep will travel from S to T in the same manner as the round trip algorithm given in Theorem B. It will deliver k_1 drums of fuel to T , which will be used on a one-way trip from T to F , plus a small amount of extra fuel for the return trip to S . On the return trip the jeep uses the k_1 drums of fuel at F to reach T , and from T uses the fuel at each depot to reach the next depot eventually returning to S . These ideas are illustrated in Figures 1 and 2 for $5\frac{3}{4}$ drums at S and $2\frac{2}{3}$ drums at F . In Figure 1 at each \mathcal{D}_i the pair (r, s) denotes r units of fuel delivered to \mathcal{D}_i for use to the right of \mathcal{D}_i and s units delivered to \mathcal{D}_i to be used on the return trip from S . In Figure 2 the number at each \mathcal{D}_i is the amount of fuel used on the return trip to the left of \mathcal{D}_i .

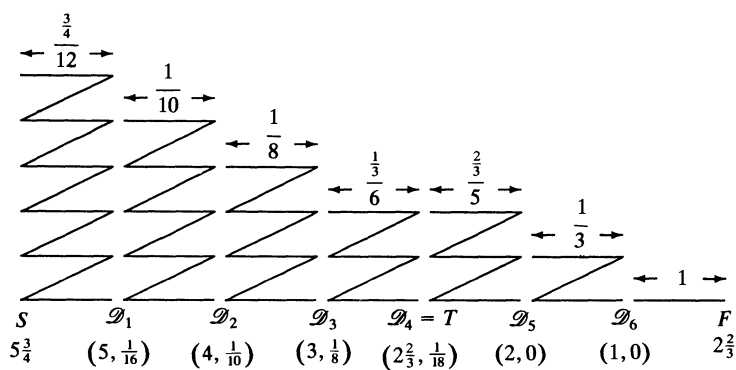


Figure 1. Outbound Trip from S to F

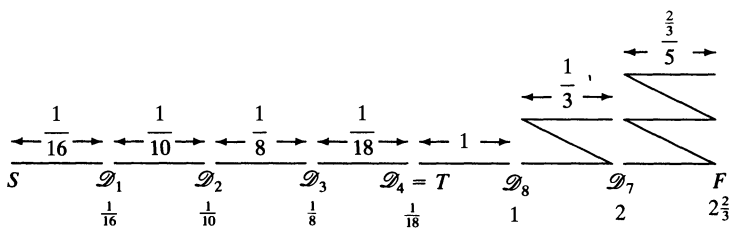


Figure 2. Return Trip from S to F

Before stating and proving Theorem 1 we make a definition which for $m = 5$, $g = \frac{3}{4}$, $k = 2$, $f = \frac{2}{3}$ gives the distance shown in Figure 1 (as well as in Figure 2). For m, k integers, $0 \leq f$, $g < 1$, and $m + g \geq k + f$ we define

$$d(m + g, k + f) = D_2 + \frac{f}{2k + 1} + \frac{1}{2k - 1} + \cdots + \frac{1}{3} + 1. \tag{1}$$

Theorem 1. Suppose there are $m + g$ drums of fuel at S and $k + f$ drums at F , where m is a positive integer, k is a non-negative integer, $m + g \geq k + f$, and $0 \leq f, g < 1$. Then the maximum distance between S and F is $d(m + g, k + f)$.

Before proving the theorem we make some definitions and prove a lemma. A *feasible solution* to a jeep problem is any trip which obeys the rules of the problem. The *value* of the feasible solution is the distance from S to F , and an *optimal solution* is any feasible solution with maximum value.

Lemma 1. For any feasible solution \mathcal{FS} that entails S -fuel arriving by jeep at point F , there is a new feasible solution \mathcal{FS}' at least as large in which no S -fuel arrives at F .

Proof of Lemma 1. Suppose S -fuel arrives at F . Without loss of generality we may assume that when S -fuel first arrives at F , say at time t_0 , there is F -fuel still there. After time t_0 each time the jeep leaves F let the percentage of S -fuel in the tank equal the percentage of S -fuel at F . Let M be the minimum amount of F -fuel which leaves F after time t_0 . Let $F' = \max\{F - M, \text{the return trip turnaround point closest to } F, \text{ return trip fuel dump closest to } F \text{ but different from } F\}$.

In the new solution \mathcal{FS}' the jeep does exactly the same movement as before, but takes no S -fuel to F , instead S -fuel is left at F' . On the return trip S -fuel which was previously taken on at F is added at F' . This proves Lemma 1.

Proof of Theorem 1. By Lemma 1 we may restrict our attention to feasible solutions in which no fuel arrives at F . We define the return trip as all travel after the jeep's first arrival at F . Let T be the point in $[S, F]$ farthest from F which can be reached from F using only F -fuel. By Theorem A,

$$\text{dist}(T, F) = 1 + \frac{1}{3} + \cdots + \frac{1}{2k-1} + \frac{f}{2k+1}.$$

We use y to denote this distance.

We give now another lemma.

Lemma 2. Suppose there is a feasible solution \mathcal{FS} in which S -fuel is used to the right of T on the return trip. Then there is a new feasible solution \mathcal{FS}' at least as large in which no S -fuel is used to the right of T on the return trip.

Proof of Lemma 2. Change \mathcal{FS} to \mathcal{FS}' as follows. On the outbound trip the travel is exactly the same as in \mathcal{FS} except any S -fuel stored in $(T, F]$ for use on the return is instead stored at T . In the return trip of \mathcal{FS}' the first travel is that required to cover the distance y from F and T . In traveling from T back to S the jeep travels as it did in \mathcal{FS} taking on fuel at T for travel to the left. This is possible because in \mathcal{FS}' we have at least as much fuel available at T as in \mathcal{FS} and the same fuel between S and T as in \mathcal{FS} . This completes the proof of Lemma 2.

By Lemma 2 we may consider only feasible solutions in which $\text{dist}(T, F) = y$ and $k + f$ units of S -fuel must arrive at T for the outbound journey. Hence we only need to maximize $\text{dist}(S, T)$, given that $k + f$ units of fuel must arrive at T for the outbound trip, and only $m + g$ units are available at S . An optimal distance and algorithm are given in Theorem B.

This completes the proof of Theorem 1.

Theorem 2 gives the maximum length desert which can be crossed using x drums of fuel divided between S and F . It shows that $k = \left\lfloor ((x+1)/2)^{1/2} \right\rfloor$ drums should be placed at F , and thus the distance from T to F will be the maximum one way distance D_1 , as given in Theorem A, that can be traveled using k drums of fuel.

Table 1 gives the number of drums at F for various values of x .

TABLE 1

x (total number of drums)	[2, 7)	[7, 17)	[17, 31)	[31, 49)	[49, 71)	[71, 97)
k (number of drums at F)	1	2	3	4	5	6

In order to prove Theorem 2 it is convenient to first prove another lemma, which shows that we can assume there are an integer number of drums at F .

Lemma 3. *Let m and k be integers, $0 \leq f, g < 1$, $m + g \geq k + f$, and $m + g + k + f \geq 2$. Then (i) $d(m + g, k + f) \leq d(m + f + g, k)$, or (ii) $d(m + g, k + f) \leq d(m + f + g - 1, k + 1)$.*

(Note: If $m + f + g - 1 < k + 1$, the right hand side of (ii) is undefined. But in that case, we will show that (i) holds.)

Proof: Suppose first that $f + g < 1$. Then we have

$$d(m + f + g, k) - d(m + g, k + f) = f \left(\frac{1}{2m + 2} - \frac{1}{2k + 1} + \frac{1}{2k + 2} \right), \tag{2}$$

which is positive when $m = k$ or $k + 1$. Now let $m \geq k + 2$. Then we also have

$$\begin{aligned} & d(m + f + g - 1, k + 1) - d(m + g, k + f) \\ &= -(1 - f) \left[\frac{1}{2m} - \frac{1}{2k + 1} + \frac{1}{2k + 2} \right] + g \left[\frac{1}{2m} - \frac{1}{2m + 2} \right]. \tag{3} \end{aligned}$$

Since

$$\left[\frac{1}{2m} - \frac{1}{2m + 2} \right] > 0,$$

the right side of either (2) or (3) will be nonnegative unless

$$\frac{1}{2m + 2} < \frac{1}{2k + 1} - \frac{1}{2k + 2} < \frac{1}{2m},$$

or equivalently

$$\frac{1}{2m + 2} < \frac{1}{(2k + 1)(2k + 2)} < \frac{1}{2m}.$$

But this forces $(2k + 1)(2k + 2)$, which is an even integer, to be $2m + 1$. This impossibility completes the proof when $f + g < 1$.

Suppose next that $1 \leq f + g < 2$. Then we have

$$d(m + f + g, k) - d(m + g, k + f) = f \left[\frac{1}{2m + 4} - \frac{1}{2k + 1} + \frac{1}{2k + 2} \right] + (1 - g) \left[\frac{1}{2m + 2} - \frac{1}{2m + 4} \right]. \quad (4)$$

The right hand side of (4) is positive when $m = k$. Now let $m \geq k + 1$. Then we also have

$$d(m + f + g - 1, k + 1) - d(m + g, k + f) = -(1 - f) \left[\frac{1}{2m + 2} - \frac{1}{2k + 1} + \frac{1}{2k + 2} \right]. \quad (5)$$

Now an argument precisely analogous to the one above shows that the right side of either (4) or (5) is nonnegative, which completes the proof of Lemma 3.

Theorem 2. *Given $x \geq 2$ drums of fuel divided between depots at each end of the desert, the maximum distance which can be crossed in a round trip is $d(x - k, k)$ where $k = \lfloor ((x + 1)/2)^{1/2} \rfloor$. Furthermore the algorithm for achieving this distance is given in the proofs of Theorems A and B.*

Proof: By Lemma 3 we may restrict our attention to feasible solutions with an integer number of drums at F . It suffices to show that

$$\text{if } t < k, \text{ then } d(x - t - 1, t + 1) \geq d(x - t, t) \quad (6)$$

and

$$\text{if } t > k, \text{ then } d(x - t + 1, t - 1) \geq d(x - t, t) \quad (7)$$

hold for positive integer $t \leq x/2$.

In the remainder of this proof let $m = \lfloor x \rfloor - t$ and $f = x - \lfloor x \rfloor$. In order to verify (6) we have:

$$\begin{aligned} d(x - t - 1, t + 1) - d(x - t, t) &= d(m - 1 + f, t + 1) - d(m + f, t) \\ &= f \left(\frac{1}{2m} - \frac{1}{2m + 2} \right) - \left(\frac{1}{2m} - \frac{1}{2t + 1} + \frac{1}{2t + 2} \right). \end{aligned} \quad (8)$$

Since $t < k = \lfloor ((x + 1)/2)^{1/2} \rfloor$, we have $t + 1 \leq ((x + 1)/2)^{1/2}$ or $\lfloor x \rfloor \geq 2(t + 1)^2 - 1$. Thus $2m = 2(\lfloor x \rfloor - t) \geq 2((2(t + 1)^2 - 1) - t) = (2t + 1)(2t + 2)$ or

$$\frac{1}{2m} - \frac{1}{2t + 1} + \frac{1}{2t + 2} = \frac{1}{2m} - \frac{1}{(2t + 1)(2t + 2)} \leq 0.$$

This and the fact that

$$\frac{1}{2m} - \frac{1}{2m + 2} > 0$$

imply that the right side of (8) is nonnegative, as required.

In order to verify (7) we have

$$\begin{aligned}
 & d(x-t+1, t-1) - d(x-t, t) \\
 &= d(m+1+f, t-1) - d(m+f, t) \\
 &\geq -f \left(\frac{1}{2m+2} - \frac{1}{2m+4} \right) + \left(\frac{1}{2m+2} - \frac{1}{2t-1} + \frac{1}{2t} \right) \\
 &= \frac{-2f}{(2m+2)(2m+4)} + \frac{1}{2m+2} - \frac{1}{(2t-1)2t} \\
 &\geq \frac{-2f}{(2m+2)(2m+2f+2)} + \frac{1}{2m+2} - \frac{1}{(2t-1)2t}. \tag{9}
 \end{aligned}$$

Since integer $t > k = \lfloor ((x+1)/2)^{1/2} \rfloor$, we have that $t \geq ((x+1)/2)^{1/2}$ or $x \leq 2t^2 - 1$. Thus $2(m+f) + 2 = 2(x-t) + 2 \leq 2((2t^2-1)-t) + 2 = (2t-1)2t$, and so

$$\frac{1}{2m+2f+2} - \frac{1}{(2t-1)(2t)} \geq 0.$$

Adding and subtracting $1/(2m+2)$ gives

$$\frac{-2f}{(2m+2)(2m+2f+2)} + \frac{1}{2m+2} - \frac{1}{(2t-1)2t} \geq 0.$$

That is, the right side of (9) is nonnegative as required.

This completes the proof of Theorem 2.

Theorem 2 can be viewed as providing the solution to the equivalent problem: Given unlimited fuel at each end of a desert of fixed length, minimize the amount of fuel required for a round trip across the desert. Since our solution gives a distance roughly one half of the harmonic number $H_{\lfloor x \rfloor - k}$, any length desert can be crossed given the availability of sufficient fuel. It is also interesting to note that although our solution is better than placing half of the fuel at each end of the desert, the difference between the two solutions is not great. In fact it is bounded above by a small constant. It is easy to show, using common identities and estimates for harmonic numbers that this difference is always less than $1 + \ln 2$. Finally, we observe that the number of intermediate fuel depots for our optimal solution is $\lceil x \rceil - 2$.

We next consider the problem of finding the position of two depots so that a desert of maximum length can be crossed on a round trip with n drums of fuel, n an integer, distributed between the two depots. Obviously one depot, B_1 , must be at the start; otherwise the jeep cannot move. At the second depot, B_2 , suppose that we have k units of fuel, where k is not necessarily an integer. Let r be the amount of fuel at B_2 which is used on the return trip and $t = k - r$ be the amount which is used to continue the trip across the desert. Without loss of generality we can assume that in an optimal solution on the return trip the jeep arrives at B_2 without fuel, for that fuel could be included in the k units stored at B_2 where k need not be an integer. Let $s + f_1$ be the amount of fuel at S , let $r + f_2$ be the amount of fuel at B_2 which is used on the return trip, and $t + f_3$ be the amount at B_2 used to continue the outbound trip where s, r, t are integers and $0 \leq f_i \leq 1$ for $i = 1, 2, 3$ such that $s + f_1 + r + f_2 + t + f_3 = n$. The maximum

TABLE 2. Optimal partitions and distance for various values of n .

n	s	r	t	Distance from S to B_2	Desert Length
3	1	1	1	1	$1\frac{1}{2}$
4	1	1	2	1	$1\frac{3}{4}$
5	2	1	2	$1\frac{1}{4}$	2
6	2	1	3	$1\frac{1}{4}$	$2\frac{1}{6}$
7	3	1	3	$1\frac{5}{12}$	$2\frac{1}{3}$
8	3	1	4	$1\frac{5}{12}$	$2\frac{11}{24}$
9	4	1	4	$1\frac{13}{24}$	$2\frac{7}{12}$
10	4	1	5	$1\frac{13}{24}$	$2\frac{41}{60}$
11	5	1	5	$1\frac{77}{120}$	$2\frac{47}{60}$
12	5	2	5	$1\frac{29}{40}$	$2\frac{13}{15}$
30	14	2	14	$\frac{1}{2}H_{14} + \frac{7}{12}$	$H_{14} + \frac{7}{12}$
31	14	3	14	$\frac{1}{2}H_{14} + \frac{37}{60}$	$H_{14} + \frac{37}{60}$

width desert which can be crossed via a round trip is:

$$\begin{aligned} \frac{f_1}{2s+2} + \frac{1}{2s} + \frac{1}{2s-2} + \cdots + \frac{1}{2r+4} + \frac{1-f_2}{2r+2} + \frac{f_2}{2r+1} + \frac{1}{2r-1} \\ + \cdots + \frac{1}{3} + 1 + \frac{f_3}{2t+2} + \frac{1}{2t} + \frac{1}{2t-2} + \cdots + \frac{1}{4} + \frac{1}{2}. \end{aligned} \quad (10)$$

This follows immediately from the fact that we can consider the problem as a round trip from B_1 to B_2 and a round trip onward from B_2 .

Thus we need to decide how to partition n drums of fuel into 3 parts so as to maximize (10). It is easy to show, similar to Lemma 3, that for any feasible solution with value given by (10) there is a feasible solution at least as large where each f_i is an integer. We state without proof Theorem 3 which gives, for arbitrary n , an optimal partition, the maximum desert length, and an optimal position for depot B_2 . Table 2, in which H_n denotes the n th harmonic number, shows these optimal partitions and distances for some sample values of n . We also observe that for an optimal solution, B_2 is placed slightly more than half way across the desert from $B_1 = S$. In order to state Theorem 3 we make two definitions:

First define

$$\begin{aligned} d(s, r) = \frac{1}{2s} + \frac{1}{2s-2} + \cdots + \frac{1}{2r+2} + \frac{1}{2r-1} + \cdots + \frac{1}{3} \\ + 1 + \frac{1}{2t} + \cdots + \frac{1}{4} + \frac{1}{2}. \end{aligned}$$

Second given n, t , positive integers with $n > t$ we define:

$$\begin{aligned} D(n, t) = \frac{1}{2s} + \frac{1}{2s-2} + \cdots + \frac{1}{2r+2} + \frac{1}{2r-1} + \cdots + \frac{1}{3} \\ + 1 + \frac{1}{2t} + \frac{1}{2t-2} + \cdots + \frac{1}{4} + \frac{1}{2}, \end{aligned}$$

where $s + r = n - t$ and $r = \lfloor ((n - t + 1)/2)^{1/2} \rfloor$.

Theorem 3. *The maximum width desert which can be crossed by a round trip using $n \geq 3$ drums of fuel is $D(n, t)$, where r is given above and $s = \lfloor (n - r)/2 \rfloor$, $t = \lfloor n - r/2 \rfloor$. The depots B_1 and B_2 are located at S and at distance $d(s, r)$ from S respectively.*

Observe that the definition of r is similar to that of k in the previous problem. The values of r , s , and t are defined by the above system of four equations, a solution to which can be found in any one of several ways. One easy way is to find r by Theorem 4 below.

The proof of Theorem 3 is similar to, but more complicated than, that of Theorem 2. Any interested reader may obtain a copy of the proof from the third author.

Theorem 4. *If n is a positive integer and m is the least positive integer such that $4m^2 + 7m \geq n$, then $r = m$.*

Proof: It suffices to show that for $t = \lfloor (n - m/2) \rfloor$, if $n = 4m^2 + 7m$ or $n = 4(m - 1)^2 + 7(m - 1) + 1$, then $m = \left\lfloor \left(\frac{(n - t + 1)/2}{2} \right)^{1/2} \right\rfloor$. Let $n = 4m^2 + 7m$, then $t = 2m^2 + 3m$ and

$$\left\lfloor \left(\frac{n - t + 1}{2} \right)^{1/2} \right\rfloor = \left\lfloor \left(m^2 + 2m + \frac{1}{2} \right)^{1/2} \right\rfloor = m.$$

Let $n = 4(m - 1)^2 + 7(m - 1) + 1 = 4m^2 - m - 2$. Then $t = 2m^2 - m - 1$ and

$$\left\lfloor \left(\frac{n - t + 1}{2} \right)^{1/2} \right\rfloor = \left\lfloor \left(\frac{2m^2}{2} \right)^{1/2} \right\rfloor = m.$$

This proves Theorem 4.

Finally from Theorem 4, it follows that $r = \left\lfloor (-7 + (49 + 16n)^{1/2})/8 \right\rfloor$.

4. DEWDNEY'S PROBLEM. Suppose a jeep, which achieves 10 miles per gallon of fuel, can carry one 50 gallon drum of fuel in addition to at most 10 gallons in its tank. Dewdney [Dew] asked for an algorithm which maximizes the one-way distance the jeep can attain using n drums available at the start.

It may appear that this problem is the same as, or at least very similar to, the one-way problem in Section 1. Dewdney's problem, however, is somewhat more subtle. Change the units so that the jeep travels 1 unit of distance on one tank of fuel. One drum holds 5 tankfuls of fuel, and the jeep can carry one drum in addition to the fuel in its tank. Fuel can be stored in drums only. Thus at most $\frac{5}{6}$ of the capacity of the jeep can be stored, whereas in Theorem A any fraction of the jeep's capacity can be stored. Theorem A gives an upper bound for the Dewdney problem once an appropriate change of units is made. A somewhat complicated optimal algorithm for the Dewdney problem is given in [Jac]. This algorithm is optimal for all n , but attains the Theorem A bound only for small n .

A friend, Ken Maddex, misunderstood a discussion with one of the authors regarding Dewdney's problem. The Maddex problem: Given an unlimited fuel supply, but only n drums for carrying fuel and only one jeep of the Dewdney kind, maximize the distance into the desert that the jeep can attain. Of course, there are both one way and roundtrip variations of this problem. As with other jeep problems finding a travel algorithm is easy but deciding optimality is apparently not easy.

Finally we note that Brauer and Brauer [Bra] considered a problem similar to Dewdney with their jeep able to carry 1 drum and its tank able to hold 1 drum. They also added the constraint that the tank could refill only when it was empty. They developed a number of algorithms but did not prove any of them optimal except for very small n .

The authors would like to thank Professor David Gale for many helpful suggestions which improved the exposition of this paper.

REFERENCES

[Alw] G. C. Alway, Crossing the desert, *Math. Gazette* 41 (1957), 209.
 [Brä] U. Brauer and W. Brauer, A new approach to the jeep problem, *Bulletin of EATCS*, June 1989, 145–154.
 [Dew] A. K. Dewdney, Computer Recreations, *Scientific American*, June 1987, 128–131.
 [Fin] N. J. Fine, The jeep problem, *Amer. Math. Monthly* 54 (1947), 24–31.
 [Gal] D. Gale, The jeep once more or jeepers by the dozen, *Amer. Math. Monthly* 77 (1970), 493–501.
 [Jac] B. Jackson, J. Mitchem, and E. Schmeichel, A solution to Dewdney’s jeep problem, *Proc. 7th International Conference in Graph Theory, Combinatorics, Algorithms and Applications*, Kalamazoo, 1992 (to appear).
 [Niv] I. Niven, Maxima and Minima Without Calculus, MAA, Washington, D.C., 1981, 204–210.
 [Phi] C. G. Phipps, The jeep problem, a more general solution, *Amer. Math. Monthly* 54 (1947), 458–462.

Hausrath:
 Department of Mathematics
 Boise State University
 Boise, ID 83725
 hausrath@math.idbsu.edu

Jackson/Mitchem/Schmeichel:
 Dept. of Mathematics and Computer Science
 San Jose State University
 San Jose, CA 95192
 jackson@sjsumcs.sjsu.edu
 mitchem@sjsumcs.sjsu.edu
 schmeich@sjsumcs.sjsu.edu

**From *The Autobiography of Malcolm X*,
 Random House, Toronto, 1992, p. 29.**

I’m sorry to say that the subject I most
 disliked was mathematics. I have thought
 about it. I think the reason was that
 mathematics leaves no room for argument.
 If you made a mistake, that was all there
 was to it.

—Malcolm X

Contributed by Nicholas Buck
 College of New Caledonia
 Main Campus 3330, 22nd Avenue
 Prince George, British Columbia
 V2N 1P8 CANADA

Count-Wheels: A Mathematical Problem Arising in Horology

Steven H. Weintraub

In this article I will describe a brilliant invention made by an unknown medieval clockmaker. In inventing this device, the “subsidiary count-wheel”, this clockmaker solved a special case of an interesting mathematical problem, whose general solution I will present here.

1. HISTORICAL BACKGROUND. The English word “clock” derives from the Latin word “clocca”, which means bell. (Compare the German word for bell, Glocke.) This etymology, surprising at first, is actually quite logical. The first mechanical clocks were so-called “tower clocks”, mounted in church towers and the towers of other tall buildings. While the very first clocks solely kept time, as soon as striking clocks were invented, sounding the hours (by ringing a bell) became the main function of clocks. To see why this is so, let us put ourselves in medieval Europe.

For peasant farmers, there is little need to tell time. Their lives are governed by the position of the sun in the sky. Only with the development of trade did this become important. Consider two merchants. In order to transact business, they must arrange a time (and place) to meet. If they agree to meet at 10:00, say, they must know when 10:00 is. To find the time they consult the town clock, and in order to do so, they must be within its range. To maximize the range, the clock is located in a tower, but the range is still limited. However, once a chiming mechanism is installed, the range is vastly increased, for, as we all know, a clock can be heard to strike the hours over a much wider area than it can be seen from.

Thus we can see that, originally, it was more important for clocks to announce the time than to keep it. Indeed, it is more accurate to think of early clocks as defining time rather than measuring it, and medieval clockmakers made much greater progress in developing striking trains, the mechanism that chimes the hours, than in developing going trains, the mechanism that keeps the time. (Subsequently, clockmaking became a race to develop more and more accurate going trains.)

The very first striking clocks simply struck once each hour. The first striking clock which actually counted the hours was installed in campanile of the church of S. Gottardo in Milan in 1336. This clock counted the hours 1, 2, 3, . . . , 24 (note the use of the 24-hour clock). Thus part of the mechanism for this clock was a “count-wheel” to count the hours.

This count-wheel is the large wheel schematically illustrated in Figure I, reprinted from [1, p. 26]. (Although the smaller wheel turns out to be our principal object of interest, we shall ignore it for the moment.) This wheel was mounted on the same axis as a wheel with (regularly spaced) gear teeth, which actually drove the striking mechanism. This system operated as follows: At midnight the wheel is

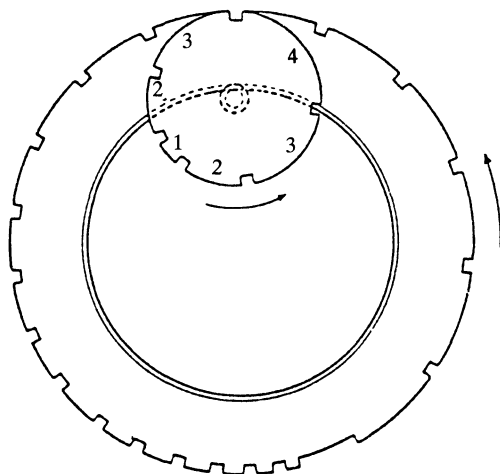


Figure I. A count wheel with subsidiary count wheel. The subsidiary wheel is mounted on the arbor that drives the principal count wheel. The number of strokes indexed on the ridges of the subsidiary wheel are marked. Indexing (on both wheels by the same indexing arm) takes place at the top; the position is after 18 hours has been struck

positioned with a lever, which prevents it from turning, in the notch at the top. At 1:00 the lever is raised, and the wheel rotates counter-clockwise until the lever falls into the next notch, stopping it. The distance between notches is such that one gear tooth engages the mechanism, causing the clock to strike once. At 2:00 a similar process occurs, but now the distance between notches is such that two gear teeth engage the mechanism, causing the clock to strike twice, etc. In the illustration the count-wheel is positioned ready to strike 19:00. (The observant reader will notice that notches for hours 1 through 4 are not individually cut, but rather there is a single long notch covering these positions. This omission will be explained below.) This quite straightforward method (which does not raise any mathematical questions) had, unfortunately, one severe drawback: it was very unreliable. To manufacture a count-wheel required making a wheel with 24 irregularly spaced notches and another with $1 + 2 + \dots + 24 = 300$ regularly spaced teeth, and this was a difficult task for 14th century metalworkers. If the count-wheel were not made precisely enough, it might, for example, stop after engaging one gear too few, ruining the striking pattern. Even a count-wheel which initially worked properly might eventually fail due to the notches or teeth wearing down in service. Finally, there was a rapid development in the miniaturization of clocks, in which table clocks became widespread, and this exacerbated the problem, for as hard as it was to construct large count-wheels for tower clocks, it was much harder to construct fine count-wheels for table clocks.

This problem was solved by an unknown genius, who came up with the idea of supplementing the original, or primary, count-wheel, with a second, or subsidiary, count-wheel. (The method of operation of this subsidiary count-wheel is our main interest here, but so as not to interrupt the historical discussion, I will defer describing it.) Not only don't we know who invented this, we don't know when, either, only that it was already known in 1475.

The subsidiary count-wheel was in use until well into the 16th century, but in the course of the 17th century this construction ceased to be used. (The most recently dated known example of a clock with a subsidiary count-wheel is from

1568.) The reason for this is two-fold. First, metal-working techniques had advanced, and second, and most decisively, the use of the 12-hour clock (i.e., counting the hours in a day $1, \dots, 12, 1, \dots, 12$ instead of $1, \dots, 24$) had won out. A primary count-wheel for a 12-hour clock requires 12 irregularly spaced notches (rather than 24) to control a wheel with $1 + 2 + 3 + \dots + 12 = 78$ (rather than 300) teeth, and these smaller numbers were within the capacity of then-current manufacturing techniques.

Our earliest information on subsidiary count-wheels comes from Frater Paulus Almanus (Brother Paul the German). Frater Almanus was a German monk who made a pilgrimage to Rome in 1475. He stayed in Rome, supporting himself by opening a shop where he bought and sold clocks, and by maintaining and repairing clocks for various Church dignitaries. To help himself out, he kept a notebook in which he recorded various noteworthy features of clock mechanisms, and among these was the subsidiary count-wheel. He compiled this notebook (written in Latin, of course) between 1475 and 1485, and it found its way to a library in Augsburg, where it lay undiscovered until the 20th century. It was translated and deciphered in stages, and John H. Leopold and Phillip G. Coole were the first to uncover and understand the section on count-wheels (among others). This manuscript is published in English translation, with commentary and explanation, in [1].

2. OPERATION OF THE SUBSIDIARY COUNT-WHEEL. The subsidiary count-wheel is the small wheel illustrated in Figure I. It is mounted so that its top and the top of the primary count-wheel are adjacent. The lever, whose fall stopped the rotation of the primary count-wheel, was made twice as wide, so that it could fall only when both the primary and subsidiary count-wheels had rotated into a position in which they both had notches at the top. In this way the subsidiary count-wheel acted as a safety device, preventing the fall of the lever unless the right number of chimes had been sounded.

Let us closely examine the operation of the subsidiary count-wheel. Note that it is irregularly divided, with notches spaced to allow turning through 1, 2, 3, 4, 3, and 2 chiming gear teeth. At midnight it is positioned in the slot to the left of the 1. At 1:00 the lever is raised, and the wheel rotates counter-clockwise 1 notch before the lever drops. At 2:00 it rotates 2 notches; at 3:00 it rotates 3; at 4:00 it rotates 4. At 5:00 it rotates $3 + 2 = 5$. (It is prevented from stopping after 3 chimes because the primary count-wheel is not in position.) Then 6:00 is counted by $1 + 2 + 3$, and 7:00 by $4 + 3$. (Here we see why we could not simply count 5:00 by 5; a break is needed to enable counting 7:00.) Then 8:00 is counted by $2 + 1 + 2 + 3$, 9:00 by $4 + 3 + 2$, 10:00 by $1 + 2 + 3 + 4$, etc. Finally, after counting 24:00, the subsidiary count-wheel has returned to its starting position, having made $20 = 300/15$ complete revolutions, ready to count the hours in the next day. (This also explains why there were not individual notches cut on the primary count-wheel for the hours 1:00 to 4:00. The clock merely relied on the more reliable subsidiary count-wheel to count these.)

By way of further clarification, let us consider Figure I, where the clock is poised to strike 19:00. When that time arrives, the count-wheels rotate in the indicated direction. The primary count-wheel (the larger one) rotates one notch, which indexes 19 strokes. The secondary count-wheel (the smaller one) rotates through the regions 4, 3, 2, 1, 2, 3, 4, indexing a total of 19 strokes as well. In doing so it makes (in this case) somewhat more than a complete revolution; to be precise, it stops when it has turned through a complete revolution plus one additional notch. At 20:00 a similar process occurs. The primary count-wheel rotates one

notch, which indexes 20 strokes. The secondary count-wheel rotates through the regions 3, 2, 1, 2, 3, 4, 3, 2, indexing a total of 20 strokes as well. In doing so it makes (in this case) somewhat more than a complete revolution; to be precise, it stops when it has turned through a complete revolution plus two additional notches. At 21:00 a similar process occurs, and so on.

Actually, there were several subsidiary count-wheels in use, but from a horological standpoint, this was the optimal one. Note that in any subsidiary count-wheel, the total number of gears counted in a complete revolution must be a divisor of 300, the total number of gears counted by the primary count-wheel in its complete revolution (in a day), so that the subsidiary count-wheel will be correctly positioned at the start of a new day. Also, as a general rule, there is more tolerance in making a rapidly rotating wheel than a slowly rotating one, so this number should be small. Finally, since a primary count-wheel was most likely to be in error by one gear-width, the subsidiary count-wheel should have as small a proportion of one gear-width segments between notches as possible. Given these criteria, this design was best.

3. THE MATHEMATICS OF COUNT-WHEELS. I will be dealing exclusively with schemes modelled on the subsidiary count-wheel, so I will drop the adjective subsidiary and simply describe them as count-wheels.

I have long been interested in horology, and several years ago I revisited the British Museum clock room (a must for anyone interested in the subject). There I saw a display describing the operation of this count-wheel. I quickly realized that not only could this count-wheel count the integers 1 through 24, it could count all positive integers!

Since I am a mathematician, the general question arose in my mind. I will denote the above mentioned count-wheel by $(1, 2, 3, 4, 3, 2)$ and call it a count-wheel of weight $15 = 1 + 2 + 3 + 4 + 3 + 2$ and length 6 (since there are six numbers in it). Define a count-wheel to be a count-wheel which can count all positive integers. My question was: Does there exist a count-wheel of weight w , for every positive integer w ?

This question is actually not a very interesting one, since a moment's reflection reveals that, for any w , $(1, 1, 1, \dots, 1)$ (w entries) is a count-wheel. We could ask whether there exists a shortest (i.e., having smallest length) count-wheel of weight w , but this is again uninteresting, since the lengths of count-wheels of weight w are positive integers, so among them is a smallest one. However, we can refine this a bit further to get some questions of real mathematical interest:

Question 1. Given a positive integer w , does there exist a unique shortest count-wheel of weight w ?

Question 2. If so, how can we construct it?

Question 3. If so, what is its length?

In this paper, we will precisely define count-wheels, and answer these questions. (The answer to question 1 is yes, for every w .) We will merely give the results here. For the proofs we refer to the reader to our paper [2].

Definition 1. A sequence $A = (a_1, \dots, a_n)$ of positive integers is a count-wheel of length n and weight $w = a_1 + \dots + a_n$ if it has the following property: Let \bar{A} be the

infinite sequence $(\bar{a}_i)_{i=1, \dots} = (a_1, \dots, a_n, a_1, \dots, a_n, \dots)$. Then there exists a sequence $0 = i(0) < i(1) < i(2) < \dots$ such that for every positive integer k ,

$$\sum_{i=i(k-1)+1}^{i(k)} \bar{a}_i = k. \quad (*)$$

Definition 2. A count-wheel $B = (b_1, \dots, b_m)$ is an amalgam of a count-wheel $A = (a_1, \dots, a_n)$ if $B \neq A$ and

$$\begin{aligned} b_1 &= a_1 = 1 \\ b_2 &= a_2 + \dots + a_{i(2)} \\ b_3 &= a_{i(2)+1} + \dots + a_{i(3)} \\ &\vdots \\ b_m &= a_{i(m-1)+1} + \dots + a_n \end{aligned}$$

A count-wheel A is called *reduced* if it has no amalgams.

For example, $(1, 2, 3, 4, 3, 2)$ is an amalgam of $(1, 2, 3, 2, 2, 3, 2)$.

Our first result is:

Theorem 3. For every positive integer w , there is a unique reduced count-wheel of weight w , denoted $[w]$. Also, $[w]$ is an amalgam of any other count-wheel of weight w .

Note that this result gives a positive answer to question 1, as $[w]$, being an amalgam of every other count-wheel of weight w , is certainly the shortest count-wheel of weight w . (Granting that $(1, 2, 3, 4, 3, 2)$ is a count-wheel, our description of its operation in the previous section shows that all of the notches in it are necessary, i.e., that it is reduced.)

Now we proceed to give an algorithm for producing $[w]$, answering Question 2. In the following definition, tB denotes the t -fold repetition of B , e.g., $2(1, 2, 2) = (1, 2, 2, 1, 2, 2)$. Note that if B is a count-wheel, so is tB , for any positive integer t .

Definition 4. A count-wheel A is *primitive* if the equation $A = tB$ only has the solution $t = 1$, $B = A$.

Our second result is:

Theorem 5. (a) The reduced count-wheel $[w]$ is primitive if w is odd. (b) If $w = 2^t v$, $t > 0$, then $[w] = 2^t[v]$.

(From part (b) we see that in fact we have if and only if in part (a). Clearly, $[1] = (1)$, so from part (b) we also see that for any $t \geq 0$, $[2^t] = (1, 1, 1, \dots, 1)$.)

Note that this theorem reduces our problem to that of constructing $[w]$ for w odd.

Theorem 6. Let w be odd. The following algorithm produces a reduced count-wheel of weight w :

Index the positions on a wheel $1, \dots, w$ clockwise and place a 1 in each position. Begin with a pointer between positions w and 1, and cut a notch there.

Step 1: For $k = 1, \dots, (w-1)/2$, successively rotate the pointer k positions clockwise and cut a notch at the point where the pointer stops, if there is not one there already.

Step 2: *Begin with the empty sequence and the pointer positioned at the notch between positions w and 1. Rotate the pointer clockwise until a notch is reached, and let the next term of the sequence be the number of 1's the pointer has passed in doing so. Do this until the pointer has returned to its original position.*

Note that Step 2 is merely counting the result of Step 1, which is the heart of the algorithm. Note also the real content of Step 1. This step merely mimics counting the hours, so if it read “For $k = 1, \dots$ ” it would be obvious. But if it read so, it would not be an algorithm, for k would be ranging over the infinite set of positive integers. Of course, there are only a finite number of notches in the wheel, so after some integer the pointer keeps falling into already cut notches. However, in order to have an algorithm, we must know what that integer is, and that is precisely what the theorem tells us. (In other words, it says that a count-wheel which counts $0, 1, \dots, (w - 1)/2$, for w odd, counts forever.) Note also that this theorem is in general sharp, in that we must start at 0 and count up to $(w - 1)/2$, and not stop sooner. (In particular, this is always the case whenever w is prime, as we see from Theorem 7(b) below.) Application of this algorithm yields $[15] = (1, 2, 3, 4, 3, 2)$, recovering the result of our unknown medieval clock-maker.

Let us illustrate this algorithm with a further example. We shall compute $[35]$. To save space, however, we will not write a large wheel, but rather work “linearly”. Further, to avoid eye-strain we will work by “breaking down” 35 rather than by “building up” from a string of 35 1's. (The reader should have no trouble seeing that this procedure is logically equivalent to that of Theorem 6.) We denote the current pointer position with a slash and already existing divisions with a comma, and we drop the leading comma. Our algorithm says we must consider $k = 0, 1, 2, \dots, (35 - 1)/2 = 17$, so it will have 18 steps. They are

$$\begin{aligned} & /35 \rightarrow 1/34 \rightarrow 1, 2/32 \rightarrow 1, 2, 3/29 \rightarrow 1, 2, 3, 4/25 \\ & \rightarrow 1, 2, 3, 4, 5/20 \rightarrow 1, 2, 3, 4, 5, 6/14 \rightarrow 1, 2, 3, 4, 5, 6, 7/7 \\ & \rightarrow 1/2, 3, 4, 5, 6, 7, 7 \rightarrow 1, 2, 3, 4/5, 6, 7, 7 \\ & \rightarrow 1, 2, 3, 4, 5, 5/1, 7, 7 \rightarrow 1, 2, 3, 4, 5, 5, 1, 7, 3/4 \\ & \rightarrow 1, 2, 3, 2/2, 5, 5, 1, 7, 3, 4 \rightarrow 1, 2, 3, 2, 2, 5, 5, 1/7, 3, 4 \\ & \rightarrow /1, 2, 3, 2, 2, 5, 5, 1, 7, 3, 4 \rightarrow 1, 2, 3, 2, 2, 5/5, 1, 7, 3, 4 \\ & \rightarrow 1, 2, 3, 2, 2, 5, 5, 1, 7, 3/4 \rightarrow 1, 2, 3, 2, 2, 3/2, 5, 1, 7, 3, 4. \end{aligned}$$

Thus we obtain $[35] = (1, 2, 3, 2, 2, 3, 2, 5, 1, 7, 3, 4)$.

We now dispose of the question of the length of reduced count-wheels, answering Question 3. We let $\lambda(w)$ denote the length of $[w]$.

Theorem 7. (a) $\lambda(2^t) = 2^t$ for all $t \geq 0$.

(b) If w is odd, $\lambda(w) \leq (w + 1)/2$, with equality if and only if w is prime.

(c) If v and w are relatively prime, $\lambda(vw) = \lambda(v)\lambda(w)$.

(d) For an odd prime p , and any $t \geq 1$,

$$\begin{aligned} \lambda(p^{2^t-1}) &= \frac{p^{2^t} - 1}{2(p + 1)} + 1, \\ \lambda(p^{2^t}) &= \frac{p(p^{2^t} - 1)}{2(p + 1)} + 1. \end{aligned}$$

For the edification of the reader, we give the values of $[w]$ for various w below. To explain our notation, given that $[5] = (1, 2, 2)$, and that $[25] = (1, 2, 2, 1, 4, 1, 4, 1, 4, 1, 4)$, we shall write $[25] = ([5], 4(1, 4))$. We then have

$[3] = (1, 2)$	$[21] = (1, 2, 3, 1, 3, 3, 2, 6)$
$[5] = (1, 2, 2)$	$[23] = (1, 2, 2, 1, 3, 1, 3, 2, 5, 1, 1, 1)$
$[7] = (1, 2, 3, 1)$	$[25] = ([5], 4(1, 4))$
$[9] = ([3], 2(3))$	$[27] = (2[9], 3(3))$
$[11] = (1, 2, 1, 2, 4, 1)$	$[45] = (1, 2, 3, 4, 5, 3, 3, 7, 2, 3, 3, 9)$
$[13] = (1, 1, 1, 3, 2, 2, 3)$	$[49] = ([7], 6(1, 2, 4))$
$[15] = (1, 2, 3, 4, 3, 2)$	$[81] = ([27], 2([9], 6(3)))$
$[17] = (1, 1, 1, 1, 2, 4, 1, 4, 2)$	$[121] = (1, 2, 3, 4, 1, [11], 9(1, 2, 3, 4, 1))$
$[19] = (1, 1, 1, 3, 1, 2, 1, 5, 2, 2)$	$[125] = ([25], 2([25], 5(1, 4)))$

As we have mentioned, the proof of these, and several other results on this subject, can be found in [2]. The methods used in these proofs are elementary.

REFERENCES

1. Leopold, J. H. *The Almanus manuscript*, Hutchinson and Co., London 1961.
2. Weintraub, S. H. Count-wheels, *Ars Combinatoria* 36 (1993), 241–247.

Department of Mathematics
Louisiana State University
Baton Rouge, LA 70803-4918

From *Trotsky a Documentary*, by Francis Wyndham and David King, Penguin, London, 1972, p. 108.

Logical arguments, even if Russell turns them into mathematical formulae, are impotent against material interests. The ruling classes will let civilization perish together with mathematics rather than give up their privileges...

—*Trotsky*

Contributed by Nicholas Buck
 College of New Caledonia
 Main Campus 3330, 22nd Avenue
 Prince George, British Columbia
 V2N 1P8 CANADA

How to Teach a Class by the *Modified* Moore Method

Donald R. Chalice

1. INTRODUCTION. Following is a description of the “Modified Moore Method” that I have used successfully in classes ranging from intermediate analysis to advanced calculus to measure theory. While using this method, I have been able to cover as much material (and in a few cases more material) as in the usual lecture-style course. More importantly, with the Modified Moore Method, the students and I have covered that material in a far more enlivening, enjoyable and intellectually stimulating way.

I have used the method with average students in average classes and with exceptional students in above-average classes and have found similar success. Over the past twenty years, I have found it much superior to simply lecturing, and if you try it I hope you will too.

Places where changes have been introduced to the usual Moore method are indicated by a “*”. For further illustration see the sample of class notes in the appendix.

TEACHING A CLASS WITH THE MODIFIED MOORE METHOD

(I) The first day. (1) The class is begun by handing out a set of notes containing the material to be presented. Each section of the notes begins with a set of *definitions*.

*If your class is the students’ first exposure to “theorem proving”, then begin your notes with a list of techniques on theorem proving from Pólya’s books [6, 7]. Use also the list from Chapter 1 in Loren Larson’s text [4].

*After the definitions are listed in the notes, you need to follow these by a short section entitled “Exercises on the Definitions”. This section consists of a set of elementary examples and exercises about the meanings of the definitions and is geared to help the average student understand the definitions as applied to simple examples. (See the appendix.)

After the “Exercises on the Definitions” section in the notes, make a list of theorems and conjectures. *Most of these theorems are true but some “red-herrings” will be useful. For example, in an introductory course such as [1], some students try to prove inductively that the union of a countable collection of closed sets is closed, thus it becomes obvious that they do not understand induction, and the discussion ensuing after their presentation is helpful to them. (*In more advanced courses, you may put two, one or no “*”’s preceding various theorems to indicate their difficulty. Surprisingly, students prefer to do theorems with a “*” than without.) *Very difficult theorems that *you alone will prove* will be preceded by an “ Ω ” or “****”.

(2) Next present the class rules. The rules I present are:

(i) “You cannot talk to anyone, (not even your wife or girl-friend or husband or boy-friend) about the proofs of the theorems until they are done in class. You can

however, come to see me and talk about your proofs during office hours or by appointment.”

(ii) “You cannot look at any textbooks pertaining to the course.¹ You cannot look at any other set of class notes or solutions.”

*(iii) “A list of proved theorems will be placed on reserve in the library after they are done in class.”

*(iv) “The quarter/semester will be divided into three periods. *An exam will be given at the end of each period on the theorems that have been proved in that period. You will receive a grade at the end of each of these periods to indicate how well you are doing.”

*(v) Generally, you (the instructor) need to send three people to the board *at once* to write up their proofs. When they are finished with their write up, then send each in turn up to the board to present his or her proof to the class.

*(vi) Encourage mistakes—up to a point. “Do not be afraid to make mistakes on the board. If I were lecturing, I would only show you the ‘right’ way to do a problem or proof. But those who make mistakes are also making a contribution in that they are showing us the limits of the proof or problem. But there will be a five minute time limit on mistakes. Generally, you (the student) should not try to patch a proof at the board if a mistake is found. Rather, if nothing comes in your allotted five minutes, make a graceful exit, perhaps saying something like ‘Well, I need to think about this some more.’”

(II) The second day. *Send at least 5 students to the board at once to write up their answers to the *exercises on the definitions*. *Thus all students are writing up their solutions *at once*. *Often assign the same exercise to more than one student to get an “alternative point of view”.

(III) The third day. Students begin presenting their proofs.

*(1) Send them to the board in *waves of three* to write up their proofs together. Then after all proofs are written up, send each back to the board in turn to present their proof to the class.

(2) Try to make the class atmosphere safe for thoughtful expression, partly, by taking the point of view that we are all working to help and encourage one-another in the endeavor of the class. Encourage a helpful attitude rather than cutting one-another down. Always let a proof or attempted proof be first presented *without interruption*, then look around the room to see whether understanding prevails. When I see a glazed look on some people, I ask if they understand the proof or problem. Sometimes, especially at the beginning they may have trouble phrasing a question. *I say “You can always ask for an ‘instant replay’.” So often, proofs are “*replayed*.” During the repeat presentation, I consider myself something of a *conductor* of the class as I can see how the presentation may be made more clear, and often after this second presentation I will make suggestions on where and how to simplify a proof or problem to make it more elegant and to improve its presentation at the board. (A lecturer, unfortunately, has no “conductor” to direct him toward making a clearer presentation!)

I now present some reasons for the modifications made above *in order of importance*. The headings refer to the particular day in class.

¹Usually do not relax this condition. But in more advanced courses, if textbooks *are* allowed then I ask for a reference if the student uses one.

Rationale for the Modifications

*III-1. It is very important that for each round of proofs that you send three students to the board *together* to write up their solutions at once. This simple device alone enables the course to go nearly as fast and sometimes faster than the corresponding lecture style course. And this way up to five or six proofs may be presented per class-hour.

*I-2. Exercises on the definitions. Especially for average students, this section is indispensable and must be included to prevent much stumbling that would otherwise occur. It also helps the students warm up, especially at the beginning and is the basic *device* that allows us to use the Modified Moore-Method, in the first place, with students of average undergraduate calibre. For *exercises* be sure to send 5 to 7 students to the board at once.

*I-3(iii). *Documenting* the solutions in the library frees students (if they wish) from taking notes and allows them to pay full attention to the proof under discussion. However, many will take notes anyway. Additionally, your (the instructor's) proofs on reserve are of course, usually more concise and easier for them to learn for the exams later.

*I-3(iv). *Grading Periods*. The three grading periods are very useful in that students will perform better and more often if you give them an *evaluation deadline*. An *exam* given at the *end of each grading period* on the proofs already presented then forces them to review the material so that they will be able to proceed to the *next level*. Give the student a grade and an evaluation at the end of each period.

Thus, with notes and exams, you should find that the above method teaches the required proofs as well, if not better than a standard lecture style class, but with far greater enjoyment on your and the students' parts and with more enhancement of the students' interaction and creativity.

REQUIREMENTS OF THE MODIFIED MOORE-METHOD

(1) Class size. In general, class size should be limited to a maximum of 24, with ideal size about 14–18. (If your university is committed to quality of instruction, then class sizes will be appropriately small since student to instructor ratio is one such measure of quality.)

(2) Office hours. It is very important for the proper functioning of the class that the instructor be there at his or her office hour. (This might be viewed partly as preparation for the class.) So an equal number of office hours as class hours is appropriate.

(3) More class hours makes the class work better. E.g., 4 hours per week seems to be the best number for most undergraduate classes, such as advanced calculus and linear algebra. Graduate classes work well with the usual 3 hour class; but some, such as Lebesgue integration, will work better if 4 class hours are allotted per week.

CONCLUSION. I feel that a part (and possibly a fairly large part) of a student's education should be exposure to courses taught by this Socratic type method. Partly, student reactions to this method convince me so. For example, when I encounter former graduate students who have jobs in the "real world" and ask them what courses they felt helped them the most, often I will get the reply that one of the Modified-Moore-Method courses did because in it they had to learn to express their own ideas convincingly and forcefully to a large group.

CLOSING AND AN INVITATION. The cornerstone to this Socratic type method of teaching is the “enjoyment of Mathematics” and “class participation” and I hope that some of you reading this article would be encouraged enough to try using this method for yourself. If I can be of any assistance with questions you might have concerning the method or with notes from my class, I welcome and encourage your queries.

APPENDIX—(From [1])

Chapter II. Open Sets and Closed Sets.

PART A: Definitions.

*PART B: *Exercises to help with the definitions (samples follow).*

2.3. Draw: a neighborhood of p ; an ε -neighborhood of p ; a deleted neighborhood of p .

2.4. Tell whether the following sets are open and why. (A list of open sets to check.)

2.5–2.6. Which of the following sets are open? Which are closed? Neither? (examples) (By standard English usage some students may think “open” is opposite to “closed”.)

2.7. Which of the points p, q, r are cluster points of the following sets and why? (examples)

2.8. Using the quantifiers \forall or \exists , symbolize the definitions for each of the following:

a) O is an open set. a') O is not an open set. etc.

Prove or Disprove (List of theorems. Samples follow.)

$$7. \left(\bigcap_{i=1}^{\infty} A_i \right) \cup \left(\bigcap_{i=1}^{\infty} B_i \right) = \bigcap_{i=1}^{\infty} (A_i \cup B_i).$$

8. Prove or disprove #7 if $A_1 \supset A_2 \supset \cdots$ and $B_1 \supset B_2 \supset \cdots$ are nested.

32. The union of any collection of open sets is open.

33. The union of any collection of closed sets is closed. (They tend to try to prove this inductively.)

Chapter III. Connectedness.

PART A: Definitions.

*PART B: *Exercises on the definitions.*

3.1. Draw: a) two disjoint sets. b) two mutually separated sets. c) two disjoint sets that are not mutually separated.

3.2. Which of the sets S below are relatively open and which are relatively closed in the semicircle $\{(x, y) | x^2 + y^2 \leq 1 \text{ and } y \geq 0\}$? (list of examples), etc.

3.4. (Do 3.4 and 3.5 on the board together.) Which of the following sets A, B are mutually separated? A pictorial list of examples like: $A = N(0, 1)$, $B = \partial N(2, 1) \sim \{(2, 1), (2, -1)\}$.

3.5. Which of the sets in 3.4 are a union of two mutually separated sets? Which of the above sets are connected?

3.7. *In your own words*, what is a component of a set? Prove (samples follow):

49. Rose Leaf Theorem.

50. If p is an element of X then there exists a component of X that contains p .^{1,2}

REFERENCES

1. D. Chalice, "Mappings and Continuity, A Modified-Moore-Method Approach," available from the author at Western Washington University, (1993).
2. B. Gelbaum and J. Olmstead, *Counterexamples in Analysis*, Holden-Day, Inc., San Francisco (1964).
3. B. Knaster and K. Kuratowski, "Sur les ensembles connexes." *Fund. Math.* 2 (1921), 206–255.
4. L. Larson, *Problem Solving through Problems*, Springer-Verlag, New York (1990).
5. M. Mandelbaum, F. W. Gramlich, A. R. Anderson, *Philosophic Problems, An Introductory Book of Readings*, The Macmillan Company, New York (1960).
6. G. Polya, *How to Solve It*, Doubleday, (1957).
7. G. Polya, *Mathematical Discovery, On Understanding, Learning and Teaching Problem Solving*, vols. I, II, John Wiley and Sons, Inc., New York (1962, 1965).
8. L. A. Steen and J. A. Seebach, Jr., *Counterexamples in Topology*, Holt, Rinehart and Winston, Inc., New York (1970).

Department of Mathematics
Western Washington University
Bellingham, WA 98225
Chalice@henson.cc.wvu.edu

¹Many false proofs of this either resemble St. Thomas Aquinas' proof of the existence of the "Prime Mover" or St. Anselm's proof of the existence of "That Than Which Nothing Greater Can Be Conceived." See [5].

²There is a connected set X in the plane such that if you remove a certain point p from it, then $X \sim \{p\}$ is "totally disconnected," i.e., the only components of $X \sim \{p\}$ are single points. [2, 3] The point p is called an "explosion point". There are three disjoint connected open sets in the plane with the same boundary. [2] Such sets at first glance seem impossible to draw. Perhaps such an "undrawable" set is a counterexample to the statement!

Mathematics is not a careful march down a well-cleared highway, but a journey into a strange wilderness, where the explorers often get lost. Rigour should be a signal to the historian that the maps have been made, and the real explorers have gone elsewhere.

—W. S. Anglin

"Mathematics and History," *Mathematical Intelligencer*, 4, #4.

The Significant-Digit Phenomenon

Theodore P. Hill

It has been frequently observed that in many tables of physical constants and statistical data, the leading digit is not uniformly distributed among the digits $\{1, 2, \dots, 9\}$ as might be expected; rather the lower digits appear much more frequently than the higher ones. Perhaps even more surprising, an exact distribution for this nonuniformity of the leading digits has been generally asserted. In 1881 Simon Newcomb [9] stated that “The law of probability of the occurrence of numbers is such that the mantissae of their logarithms are equally probable,” and concluded that

$$\text{Prob (first significant digit} = d) = \log_{10}(1 + d^{-1}), \quad d = 1, 2, \dots, 9. \quad (1)$$

(For example, (1) predicts that the leading digit is 1 with probability about .301, and at the other extreme, is 9 with probability .046.)

Although Newcomb offered no statistical evidence for (1), its rediscovery by the physicist Benford [2] some fifty-seven years later was supported by empirical evidence based on frequencies of significant digits from twenty different tables including such diverse data as surface areas of 335 rivers, specific heat of thousands of chemical compounds, and square-root tables. The union of his tables comes surprisingly close to the frequencies predicted in (1), and, Newcomb’s earlier paper having been overlooked, those frequencies came to be known as *Benford’s Law*, or the *First Digit Law*. In fact, Benford’s data not only comes surprisingly close, it comes *suspiciously* close to the predicted frequencies; Diaconis and Freedman [5, p. 363] offer convincing evidence that Benford manipulated the round-off errors to obtain an even better fit. But even the unmanipulated data seems a remarkably good fit, and the “law” has become widely accepted.

CLASSICAL EXPLANATIONS. Since Benford, numerous “mathematicians, statisticians, economists, engineers, physicists and amateurs” [11, p. 521] have attempted to explain the probabilities appearing in (1) based on a variety of hypotheses. The classical explanations include: the usual number-theoretic (or Cesaro) method for assigning densities to the sets in question; continuous analogs of the Cesaro method based on integration techniques; various probabilistic urn-schemes; demonstrations based on assumptions of continuity and scale-invariance (see below); and statistical descriptive arguments. For an excellent review of these ideas, the reader is referred to Raimi [11]. (A more recent explanation of Schatte [12] gives Benford’s Law as a corollary to an “unproved” ([12, p. 452]) “hypothesis that after a sufficiently long computation in floating-point arithmetic, the occurring mantissas have a nearly logarithmic distribution.”)

All of these previous explanations suffer from two substantial shortcomings. First, the previous methods for prescribing frequencies for such sets as “first significant digit = 1” are *not unique*. Such a set does not have a natural density,

unlike the set of even numbers, say, which has density $1/2$ among the integers and density 0 among the real numbers, and in general there are many ways of assigning a number to the set “first significant digit = d ” which are consistent with natural density. The explanations mentioned above simply single out particular summation or integration techniques that yield the “correct” Benford frequencies.

The second shortcoming is that, terminology notwithstanding, the past frequency-assigning functions leading to (1) are *not probabilities*, at least not in the classical sense. The standard mathematical definition of probability is a $[0, 1]$ -valued function P on a domain of sets (called a sigma algebra) closed under complements and countable unions, which assigns 1 to the whole set and assigns measure $\sum_{n=1}^{\infty} P(A_n)$ to the set $\bigcup_{n=1}^{\infty} A_n$ if the $\{A_n\}$ are disjoint. But the methods above necessarily fail to satisfy these conditions, as will, for example, any reasonable notion of density on the natural numbers which assigns density 0 to singletons, for then $P(\mathbb{N}) = \sum P(\{n\}) = 0 \neq 1$. (This is exactly the same reason for the foundational difficulty in making rigorous sense of “pick an integer at random”; e.g., see De Finetti [4] pages 86, 98–99). For the integer-based models of Benford’s Law, this difficulty seems insurmountable, and for the above-mentioned real-number models either a precise domain for the probability in (1) was not specified by Newcomb *et al.*, or when specified was simply not the appropriate collection \mathcal{A} .

THE PROPER PROBABILITY DOMAIN. The first step toward making rigorous sense of the First-Digit Law (1) is to identify an appropriate domain for the probability. A typical set in the desired collection \mathcal{A} of subsets of \mathbb{R}^+ is the set of positive reals whose first significant digit (base 10) is 1, namely,

$$\{D_1 = 1\} := \bigcup_{n=-\infty}^{\infty} [1, 2) \cdot 10^n.$$

This set (along with its analogs from the second, and general n th-digit laws, also known to Newcomb and Benford) suggests the following natural domain \mathcal{A} for a general significant-digit law.

Definition. \mathcal{A} is the smallest collection of subsets of the positive reals which contains all sets of the form $\bigcup_{n=-\infty}^{\infty} (a, b) \cdot 10^n$, and which is closed under complements and countable unions.

The following properties of \mathcal{A} are easy to check:

every non-empty set in \mathcal{A} is infinite, with accumulation points at 0 and at $+\infty$;
 \mathcal{A} is closed under scalar multiplication, i.e. $a > 0$ and $S \in \mathcal{A} \Rightarrow aS \in \mathcal{A}$;
 \mathcal{A} is self-similar, in the sense that if $S \in \mathcal{A}$ and $k \in \mathbb{Z}$ then $10^k S = S$.

For each $i = 1, 2, \dots$, let $D_i: \mathbb{R}^+ \rightarrow \{0, 1, \dots, 9\}$ be the i th significant-digit function, for example, $D_1(\pi) = 3$, $D_2(\pi) = 1 = D_2(10\pi)$. It may easily be shown [8] that

$$D_i^{-1}(\{d\}) \in \mathcal{A} \text{ for all } i \text{ and } d,$$

and in fact, \mathcal{A} is the smallest such collection (closed under complements and countable unions) for which this is true. (In measure-theoretic terms, \mathcal{A} is the sigma-algebra generated by D_1, D_2, \dots) This shows that \mathcal{A} is precisely the correct domain for a general significant-digit probability law.

THE GENERAL SIGNIFICANT-DIGIT LAW

General Significant-Digit Law [8]. For all $k \in \mathbb{N}$, all $d_1 \in \{1, 2, \dots, 9\}$ and all $d_j \in \{0, 1, 2, \dots, 9\}$, $j = 2, \dots, k$,

$$P\left(\bigcap_{i=1}^k \{D_i = d_i\}\right) = \log_{10} \left[1 + \left(\sum_{i=1}^k d_i \cdot 10^{k-i} \right)^{-1} \right]. \quad (2)$$

Observe that this *joint* significant-digit law (2) includes the First-Digit Law (1) as a special case, as well as the other marginal significant-digit laws.

Example.

$$P((D_1, D_2, D_3) = (3, 1, 4)) = \log_{10} \left(1 + \frac{1}{314} \right) \cong .0014.$$

A perhaps surprising corollary of (2) is that

the significant digits are dependent

and not independent as one might expect. For example, from (2) it follows that the (unconditional) probability that the second digit is 2 is $\cong .109$, but the (conditional) probability the second digit is 2, *given* that the first digit is 1, is $\cong .115$. Similarly, the hundredth significant-digit is also dependent on the first few significant digits, although the dependency decreases as distance between the digits increases. It also follows easily from (2) that the distribution of the i th significant digit approaches the uniform distribution (where each digit $\{0, 1, \dots, 9\}$ occurs with frequency $\frac{1}{10}$) exponentially fast as $i \rightarrow \infty$.

What simple hypotheses lead to the General Significant-Digit Law (2)?

SCALE AND BASE-INVARIANCE. One set of hypotheses which has been popular in the past is the notion of *scale-invariance*, which corresponds to the following idea. If the first digits obey some fixed universal distributional law, then this law should be independent of the units chosen (e.g., English or metric systems). However, as Knuth pointed out (cf. Raimi [11]), *there is no scale-invariant probability measure on the Borel subsets of \mathbb{R}^+* , since then the measure of the set $(0, 1)$ must be the same as the measure of *every* interval $(0, b)$, which by countable additivity must be 0.

The problem is simply that the Borel sets (the smallest sigma-algebra containing all open intervals) are not the appropriate domain for the significant-digit probability law; using \mathcal{A} instead resolves this problem.

On \mathcal{A} , it is easily shown [8] (since the orbit of every point under irrational rotation on the circle is asymptotically uniformly distributed) that if P is scale-invariant, i.e., if $P(bS) = P(S)$ for all $b > 0$ and all $S \in \mathcal{A}$, then P satisfies (2). That is, on the correct domain \mathcal{A} ,

scale-invariance implies Benford's Law.

One possible drawback to the scale-invariance hypothesis is the special role played by the constant 1. In most tables of physical constants, the constant 1 simply does not appear, since the underlying law (say, in $f = ma$) does not necessitate definition of a constant (as opposed to $e = mc^2$). If a “complete” table of physical constants included the constant 1, perhaps that special constant would occur with

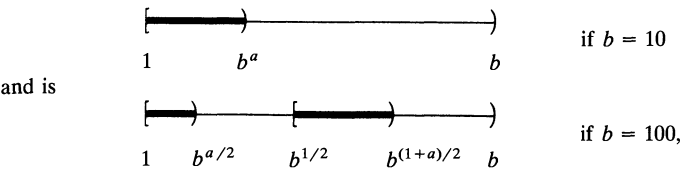
strictly positive frequency. But this would preclude scale-invariance, since then $0 < P(\{1\}) = P(\{2\}) = \dots$, contradicting the additivity of a probability.

As an alternative hypothesis, suppose that any universal significant-digit law were *base-invariant*; i.e., carried over to bases other than 10. (As pointed out in [11], all the classical arguments supporting (1) and (2) carry over *mutatis mutandis* to other bases such as 2, or 7 or 100.)

To motivate a formal definition of base-invariance, consider the set of positive numbers S with first significant digit (base 10) less than 5. Using the decimal notation D_1 as above, and letting $D_1^{(100)}$ denote the first significant digit base 100, it is easily seen that

$$S = \{1 \leq D_1 < 5\} = \{1 \leq D_1^{(100)} < 5\} \cup \{10 \leq D_1^{(100)} < 50\},$$

which says that graphically (as a subset of $[1, b)$), the *same* set S is



(where $a = \log_{10} 5$). Hence if a probability P on \mathcal{A} is “base-invariant,” the measures of these two S -representing subsets of $[1, b)$ should be the same, i.e.,

$$P([1, b^a)) = P([1, b^{a/2})) + P([b^{1/2}, b^{(1+a)/2})),$$

and similarly for higher power bases b^n . This suggests the following definition.

Definition. [8] P is *base-invariant* on \mathcal{A} if

$$P([1, 10^a]) = \sum_{k=0}^{n-1} P[10^{k/n}, 10^{(k+a)/n}) \quad \text{for all } n \in \mathbb{N} \text{ and all } a \in (0, 1).$$

Letting P_L be the logarithmic probability defined in (2) and P_0 be the degenerate probability which assigns mass 1 to the constant 1 (or formally, to the set $\bigcup_{n=-\infty}^{\infty} \{10^n\}$ in \mathcal{A}), it now follows [8] using a slightly deeper result from ergodic theory concerning invariant measures on the circle, that

$$P \text{ is base-invariant} \iff P = qP_0 + (1 - q)P_L \quad \text{for some } q \in [0, 1].$$

Corollaries are:

the logarithmic distribution (2) is the unique continuous base-invariant distribution

and

scale-invariance implies base-invariance.

(Observe that base-invariance does *not* imply scale-invariance, since P_0 is base but not scale-invariant.) Thus, if there is a universal significant-digit law and it is base-invariant, then the special constant 1 occurs with possibly positive probability q , and otherwise (with probability $1 - q$) the digits satisfy the logarithmic distribution (2).

APPLICATIONS

Computer design and analysis of roundoff errors. Hamming [6] has given applications of Benford's Law to the problem of placing the decimal (binary) point in the number system of a computer in order to minimize the number of normalization shifts after the computation of a product, to the problem of estimation of the representation error of numbers in base 2 and base 16, and to the problem of roundoff error propagation. Schatte [12] similarly concludes that the choice of a binary-power base $b = 2^r$ can be guided by the hypothesis of logarithmic distribution (cf. Benford's Law) of mantissa errors; for example, he argues that base $b = 2^3$ is optimal with respect to storage use.

Statistical Tests for "Naturalness." Varian [13] has proposed using Benford's Law as a test of "reasonableness" for data, by checking forecasts of a mathematical model as to goodness of fit to Benford's Law. He used this idea to check specific models for economic production and for forecasts of acres of land in various use, and Becker [1] used Benford's Law to check lists of failure rates to detect systematic errors. The underlying idea in these applications is that if "real life data" obeys Benford's Law, then so should good mathematical models.

Making Money in Numbers Games. In the Massachusetts Numbers Game [cf. 3], players first bet on a four-digit number of their choice, next a single four-digit number is generated randomly by an umpire, and then all players with the winning number share the (tax-reduced) pot equally. In such a situation it is obviously advantageous to identify numbers which few people choose, since all numbers are equally likely to be winners and the expected payoff for an unpopular number is thus higher than that for a number which many people have chosen. Now *if* people choose numbers from their experience, and *if* the numbers in their experience obey Benford's Law, then it makes sense to pick numbers *inversely* to Benford's Law, i.e., numbers starting with 9 or 8. Of Chernoff's [3] 33 statistically-obtained numbers in his "first system" (numbers with predicted normalized payoffs exceeding 1.0) for playing the Massachusetts Numbers Game, 16 had first significant digit 8 or 9, and only 1 has first significant digit 1 or 2. (Additional evidence that numbers "randomly" generated by people tend to start with low digits is found in Hill [7].) Since Chernoff also concluded that the public learns quickly, this suggests using inverse-Benford as an initial strategy when a new numbers game is initiated, and then quitting play soon thereafter.

Outfoxing the Internal Revenue Service. In his Ph.D. thesis, Nigrini [10] has suggested that the IRS use Benford's Law as a test for detecting fraud, such as falsification of data by a taxpayer at the time of filing his return. Nigrini's hypothesis is that true data gives a rough approximation to Benford's Law, whereas a Benford-ignorant cheater tends to concoct numbers according to some other distribution, say uniform via a standard random number generator, or more likely, a subconscious personal favorite generated mentally. Nigrini proposes that the IRS simply check for goodness-of-fit against Benford, and then audit the worst fits. This suggests that a "creative" and Benford-wise taxpayer should modify or generate his fabricated data according to a Benford-like distribution.

ACKNOWLEDGMENT. The author is grateful to Professors Bob Foley and Ron Fox for several useful suggestions and Göran Högnäs for pointing out the self-similarity of A .

REFERENCES

1. Becker, P. (1982) Patterns in listings of failure-rate and MTTF values and listings of other data. *IEEE Transactions on Reliability*, R-31, 132–134.
2. Benford, F. (1938) The law of anomalous numbers. *Proc. Amer. Phil. Soc.* 78, 551–72.
3. Chernoff, H. (1981). How to beat the Massachusetts Numbers Game. *Math. Intel.* 3, 166–172.
4. De Finetti, B. (1972) *Probability, Induction and Statistics*. Wiley, New York.
5. Diaconis, P. and Freedman, D. (1979) On rounding percentages. *J. Amer. Stat. Assoc.*, 359–64.
6. Hamming, R. (1976) On the distribution of numbers. *Bell Syst. Tech. J.* 49, 1609–25.
7. Hill, T. (1988) Random-number guessing and the first digit phenomenon. *Psychological Reports* 62, 967–71.
8. Hill, T. (1995) Base-invariance implies Benford's Law, *Proc. Amer. Math. Soc.* 123, 887–895.
9. Newcomb, S. (1881) Note on the frequency of use of the different digits in natural numbers. *Amer. J. Math* 4, 39–40.
10. Nigrini, M. (1992) The detection of income evasion through an analysis of digital distributions. Ph.D. Thesis, Department of Accounting, University of Cincinnati.
11. Raimi, R. (1976) The first digit problem. *Amer. Math. Monthly* 83, 521–38.
12. Schatte, P. (1988) On mantissa distributions in computing and Benford's Law. *J. Inf. Process. Cybern.* EIK 24, 443–455.
13. Varian, H. (1972) Benford's Law. *Amer. Statistician* 26, 65–66.

School of Mathematics
Georgia Institute of Technology
Atlanta, GA 30332
hill@math.gatech.edu

“...in the current state of analysis we may regard the discussion [of past mathematics] as tasteless, for they concern forgotten methods, which have given way to other more simple and more general. However, such discussions may yet retain some interest for those who like to follow step by step the progress of analysis, and to see how simple and genereal methods are born from particular questions and complicated and indirect procedures.”

—J. L. Lagrange

Exploring the Brachistochrone Problem

LaDawn Haws and Terry Kiser

1. MAKING THE BRACHISTOCHRONE ACCESSIBLE. In light of the attention given to a national crisis in mathematics education, concerned mathematics instructors are always looking for innovative ways to present and reinforce ideas. For a generation that grew up with fast paced MTV and special effects movies like Star Wars, the classroom may appear to be a fairly dull environment with uncompromising standards. Computer technology can help educators compete for students' attention and at the same time enhance the learning process by

- 1) bringing an added dimension—visualization—to the presentation of mathematical concepts,
- 2) giving students greater flexibility to explore and discover ideas on their own,
- 3) making more advanced topics accessible to a wider range of classes.

These learning aspects will be discussed in the context of some *Mathematica* packages for exploring the classic Brachistochrone problem and interesting variations.

The Brachistochrone Problem, to find the curve joining two points along which a frictionless bead will descend in minimal time, is typically introduced in an advanced course on the Calculus of Variations. The statement of this problem is easily understood, even for high school students, when phrased in a more familiar context as follows: “What shape should a roller coaster track have so the car will travel from a high point A to a low point B as fast as possible?” This form of the statement of the problem, however, has resulted in some unexpected and amusing responses from students who were asked to draw what they thought would be a “fast track.”

The authors have written a *Mathematica* command called **Race** that allows students to explore this problem graphically. We have developed several activities and exercises for students with a wide variety of mathematical abilities, from algebra to differential equations. The student may design a path, or several paths, and **Race** will produce a plot of the paths, their lengths, times of descent and, optionally, an animation of beads racing down the paths.

The exciting part is that **Race** enables students to experiment on their own with different shaped curves to gain intuition and formulate criteria for a “fast curve” without needing the mathematical expertise to solve the problem. Articulating criteria such as,

- the curve should start out with a steep descent to build up velocity quickly, but
- the steep part should not be “too long” or the advantage gained in increased acceleration will be lost in increased path length,

requires a good understanding of slope, velocity, acceleration, and arc-length; fairly sophisticated stuff for pre-calculus or even high school students!

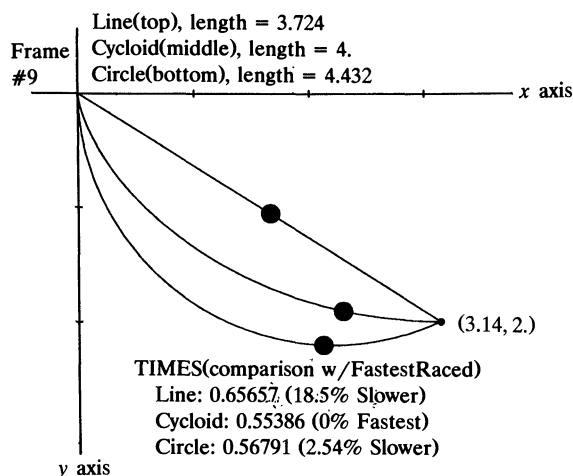


Figure 1. A *Mathematica* simulation generated by *Race*

The well known solution curve, the cycloid, makes the optimal compromise between steepness versus path length and is easily derived from the Euler-Lagrange equation. Of course, this tool is not available in most undergraduate mathematics classes, but that does not mean the underlying problem is inaccessible to these students. Instead of simply presenting the cycloid as an interesting example of a parametric curve, as is typically done in a 1st or 2nd semester calculus class, its special features can be developed—in fact, the students can discover them for themselves.

Students in a beginning differential equations course can understand the derivation and solution of the differential equation governing the Brachistochrone. They are already familiar with minimization criteria (to minimize $f(x)$, consider solutions to $f'(x) = 0$), so the Euler-Lagrange equation is not hard for them to swallow and should not be a deterrent from investigating this application.

Use of *Mathematica* as a basis for exploring the Brachistochrone problem is a prime example of how technology can allow students to go beyond standard textbook applications and address more realistic or current applications. For example, a natural question to ask concerning the Brachistochrone problem is what happens if friction or air resistance is included in the model? This is discussed later in this article. The messy calculations that are typical in many real-world applications and (up to now) made them off limits in the undergraduate classroom can be handled by the computer, with possibly some surprising results to student and instructor alike.

There are many applications that computer technology now makes accessible at all levels. One challenge to all of us as educators is to make creative use of this technology. The *Mathematica* command **Race** and accompanying packages along with a notebook of examples is available upon request. This article will conclude with some specific examples illustrating these exercises.

2. PRE-CALCULUS AND CALCULUS. An engaging way to present this topic is to begin by testing the students' intuition. We have broken classes up into groups and asked them to draw and discuss what they think is the fastest curve. In addition, we have created a *Mathematica* package that allows students to “draw”

their curve on the computer screen and generates a simulation from a speedier, scaled down version of **Race** (this has been used several times to provide a computer lab experience for 7th–9th grade students as a part of a “Math Field Day” project held annually at Chico State University). We next give a “live” demonstration of marbles racing down wooden ramps, roughly in the shape of a cycloid and a straight line. This demonstration generates a great deal of excitement, which just goes to show that the “old technology” still has its place. It also provides a concrete time scale which is necessary for qualifying what constitutes a “close” race. Even in our differential equations classes where the majority of the students deduce on their own that the marble rolling down the cycloidal ramp will beat the marble on the straight line ramp, many are surprised at the margin of victory.

The focus of the presentation should be a graphical exploration of the criteria that makes the cycloid a fast curve, not on its derivation as the fastest curve. The details concerning the cycloid can be adjusted depending on the level of the class. We use a *Mathematica* animation to present the cycloid as the curve generated by tracing a point on a rolling circle (due to Stan Wagon at Macalester College) and in a Calculus course it is appropriate to derive the parametric equations. Even after the cycloid is presented as the solution to the Brachistochrone problem, there are many interesting questions to investigate that depend on physical or graphical intuition. For example, a traditional graphing exercise for a pre-calculus class can be spiced up by asking for the fastest curve among a class of familiar functions (especially appropriate if these functions have recently been studied) but with an unknown parameter. This motivates the need to graph several examples of the function in question so they can apply their newly gained intuition on what makes a fast curve. Then, they can **Race** their graphs to check their intuition. Test your intuition on the examples below.

2.1. Finding a Fast Parabola. Find the fastest parabola that starts at the origin and ends at a given point, say, (3, 2) (in this article we will always take the positive y -axis to be oriented downward). Since there are three unknowns in a quadratic, we are free to impose one more condition. Let’s take the x -coordinate of the vertex, m , to be our unknown parameter. Below are graphs for $m = 2, 2.5, 3,$ and 3.5 . Which of these is the fastest curve and how does this value of m compare with the optimal value for m giving the fastest of all such parabolas? The cycloid ending at the point (3, 2) is included to add some perspective. The *Mathematica* generated plot of times of descent versus m for a wide range of values of m indicates that the fastest of all such parabolas is found with $m \approx 2.5$ (not the parabola with its vertex at the ending point which is a popular choice—numerical minimization verifies that $m \approx 2.494$).

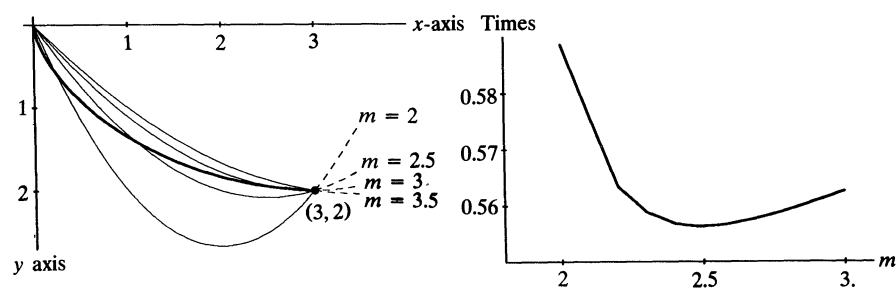


Figure 2. Finding a fast parabola

2.2. Finding the Fastest N th Root. This is a good family of curves for exploring the trade off between steepness versus path length. Below are graphs of $y = 2(x/\pi)^{1/n}$, for $n = 2, 4$, and 6 . The fastest of these curves is very competitive with the cycloid, being only slightly more than 1% slower; this would not be discernible with a live model. Can you tell which one it is?

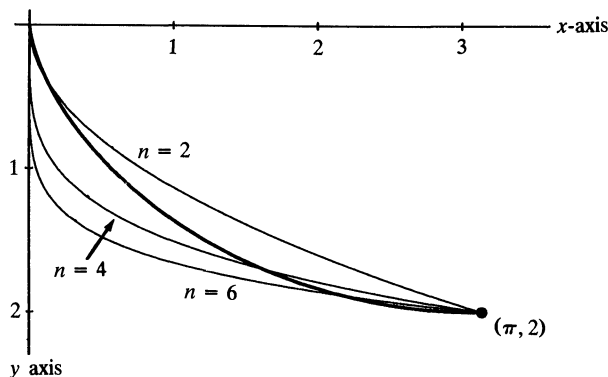


Figure 3. N th roots-Steep vs Path length

The answer is $n = 2$. This has always surprised our audiences, both faculty and students, especially when a plot of the cycloid is not included.

3. ADDING A SMALL DOSE OF REALITY. Inevitably, any discussion about the importance of computer technology to mathematics will bring up the ability to address more realistic applications. We strongly agree and yet here we are presenting an application that is only valid if it takes place in a vacuum and we ignore frictional forces! It is vital that students don't leave our classes with incorrect insight because of the setting we choose to present an application. Is this a critical issue for this problem? What happens if kinetic friction or air resistance is included in the Brachistochrone model? The *Mathematica* simulation allows the user to include a coefficient of friction to see its effect on the descent time for any curve. A typical coefficient of friction, to be denoted as μ , will be less than or equal to 0.1. A cycloid ending at the point $(\pi, 2)$ with $\mu = .1$ is approximately 3% slower than a cycloid with no friction.

Although kinetic friction has seemingly little effect on the travel time, the question still remains as to what is the effect on the shape of the fastest curve when friction is included in the model for the Brachistochrone problem. Is the cycloid still the fastest curve? If there is a new fastest curve, is it initially steeper than the cycloid lying below it or shallower and lying above it? Physical insight alone can play a key role in answering these qualitative questions. The solution when kinetic friction is included (this will be derived below for a simplified model of friction) has significantly different graphical characteristics and is an interesting generalization of the cycloid.

It is helpful to begin by motivating what graphical and physical insight suggests will be the effect of including friction. Depending on the level of the course, the solution can be derived or merely presented graphically and compared to other curves. The frictional force is assumed to be proportional to the normal component of the weight of the bead and acting in the negative tangential direction (see Figure 4). Due to the curvature of the path, the normal component of acceleration

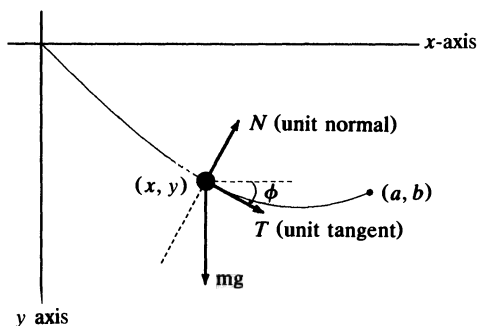


Figure 4

also contributes to the frictional force. We generally neglect this component of the friction in our initial discussions and derivations with students in a differential equations course. Students are more familiar with the weight component from studying inclined planes in physics and this is usually challenge enough. The more realistic solution, however, is presented graphically and can lead to interesting discussions as to why it differs from the cycloid or the new Brachistochrone using the simpler model of friction. This simplified model incorporates some interesting qualitative changes and has the additional advantage that the derivation of the solution is accessible in an introductory differential equations course since the equation of motion remains integrable (a more accurate treatment requires a constrained variational technique and, amazingly enough, it can still be solved in terms of elementary, albeit, messy functions—see [1]).

Before proceeding with a derivation of the solution, let's develop some qualitative insight by doing a simple physical analysis, that is, let's compare the forces with friction included versus no friction where we know the shape of the solution is a cycloid. Neglecting curvature, the magnitude of the force of friction is less at steep points on a curve, ranging from zero at a vertical tangent to the whole weight at a horizontal tangent. Since the lesson learned from the classical Brachistochrone problem, heuristically speaking, is that steepness is most important initially, this suggests that steepness will now be given more weight (versus path length) and the optimal curve, which should still have an initial vertical tangent, will be slightly steeper or below the cycloid (at least for the “beginning” portion of the curve). Since the normal component of acceleration is proportional to the square of the speed, one might expect just the opposite to be the case when it is included in the model for friction. Starting off steeper would force more curvature for the latter portion of the path when there is a greater velocity.

3.1. Derivation of the Fastest Curve With Kinetic Friction. Take the starting point to be the origin and orient the positive y -axis downward. We seek the fastest curve $y(x)$ starting at $(0, 0)$ and ending at an arbitrary point (a, b) .

If we ignore friction, then we can apply the conservation of energy, or equivalently, equate work with change in kinetic energy to obtain $v = \sqrt{2gy}$ where the velocity v is given as ds/dt . Including friction forces us to do a line integral to find the work or, alternatively, we can start with the equation of motion as follows. At a point (x, y) on the curve, the unit tangent and normal vectors, illustrated in Figure

4, can be written in terms of arc-length s as,

$$\mathbf{T} = \frac{dx}{ds}\mathbf{i} + \frac{dy}{ds}\mathbf{j} \quad \text{and} \quad \mathbf{N} = -\frac{dy}{ds}\mathbf{i} + \frac{dx}{ds}\mathbf{j}.$$

The forces of gravity and friction are given by,

$$\mathbf{F}_{gravity} = mg\mathbf{j} \quad \text{and} \quad \mathbf{F}_{friction} = -\mu(\mathbf{F}_{gravity} \cdot \mathbf{N})\mathbf{T} = -\mu mg \frac{dx}{ds}\mathbf{T}.$$

So, the components along the curve (i.e. in the direction of \mathbf{T}) are

$$\mathbf{F}_{gravity} \cdot \mathbf{T} = mg \frac{dy}{ds} \quad \text{and} \quad \mathbf{F}_{friction} \cdot \mathbf{T} = -\mu mg \frac{dx}{ds}.$$

Using these components in Newton's first law gives,

$$m \frac{dv}{dt} = mg \frac{dy}{ds} - \mu mg \frac{dx}{ds} \quad (1)$$

and substituting

$$\frac{dv}{dt} = v \frac{dv}{ds} = \frac{1}{2} \frac{d}{ds}(v^2)$$

into (1) yields, after integration w.r.t. s ,

$$\frac{1}{2}v^2 = g(y - \mu x) \quad \text{or} \quad v = \sqrt{2g(y - \mu x)}.$$

Apply the chain rule to $v = ds/dt$ and use the arc-length formula for ds/dx to solve for dx/dt as a function of x , which can be inverted to give the total time,

$$\mathbf{T}(x, y, y') = \int_a^b \sqrt{\frac{1 + (y')^2}{2g(y - \mu x)}} dx. \quad (2)$$

Since the computations become quite messy, what follows will just be an outline of the major steps. Apply the Euler-Lagrange equation,

$$\frac{d}{dx}(F_{y'}) - F_y = 0$$

where F is the integrand in equation (2) to obtain the 2nd order differential equation,

$$(1 + (y')^2)(1 + \mu y') + 2(y - \mu x)y'' = 0.$$

Through two substitutions and a partial fractions integration, this can be reduced to,

$$\frac{1 + (y')^2}{(1 + \mu y')^2} = \frac{C}{y - \mu x}, \quad (3)$$

for some non-negative constant C .

Following the lead from the classical problem, the substitution $y' = \cot(\theta/2)$ into (3) can be used to obtain a parametric solution for the optimal curve. Denoting the parametrization for the cycloid as,

$$x_c(\theta) = \rho(\theta - \sin \theta) \quad \text{and} \quad y_c(\theta) = \rho(1 - \cos \theta),$$

the new fastest curve for the “frictional” Brachistochrone problem can be given in the form,

$$\begin{aligned} x(\theta) &= x_c(\theta) + \mu\rho(1 - \cos \theta) \\ y(\theta) &= y_c(\theta) + \mu\rho(\theta + \sin \theta). \end{aligned} \tag{4}$$

The parameterizations in (4) and for the cycloid are valid for a range $0 \leq \theta \leq \theta_f$, where ρ and θ_f must be determined so the curves pass through the ending point (a, b) .

Figure 5 compares this new curve with the cycloid. Note the similar repetitive pattern with vertical tangents at even multiples of $\pi\rho$; however, the minimums do not occur at the same place. We have indicated a sloping line at which this new curve stops as opposed to making it back to the x -axis.

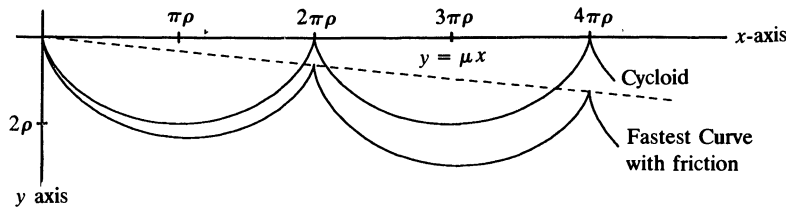


Figure 5. A generalization of the Cycloid

This line is called the “line of repose” and has the minimum slope, μ , for which the bead will begin to slide. The solution derived above is not valid for $y < \mu x$. This places a restriction on allowable ending points, which is consistent with our physical insight that, due to the loss of energy to friction, the bead can’t make it back to its original height. Given a valid ending point, there is a unique curve of this form starting at the origin with a vertical tangent. Figure 6, generated from **Race**, gives a better comparison of this “frictional” Brachistochrone versus the classic Brachistochrone, a cycloid. Both were raced with a coefficient of friction $\mu = 0.1$ and their times are compared to the cycloid without friction.

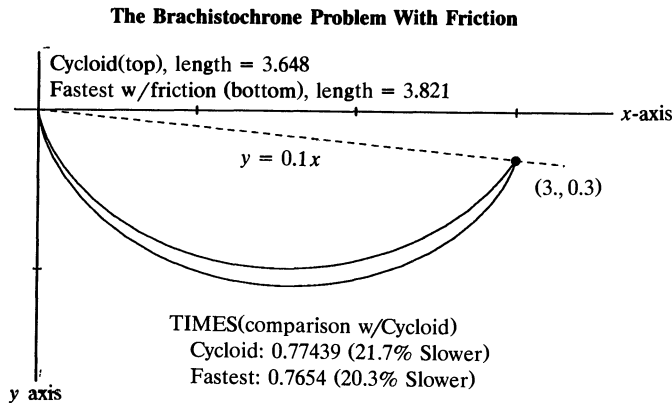


Figure 6. A race between Brachistochrones with coefficient of friction, $\mu = 0.1$

For all valid ending points, the new fastest curve lies below the cycloid throughout its entire length, and placing the ending point on the line of repose gives the greatest distinction between the two curves. With a more accurate treatment of the frictional forces, as mentioned earlier, the fastest curve lies *above* the cycloid for its entire length (see Figure 7 with $\mu = 0.1$) and the ending point must satisfy the *strict* inequality, $y < \mu x$, since the bead takes infinite time to reach the line of repose. Consistent with our earlier intuition that the normal component of acceleration puts a higher penalty on curvature, Figure 7 indicates that as the coefficient of friction increases the more realistic frictional Brachistochrone will approach a straight line.

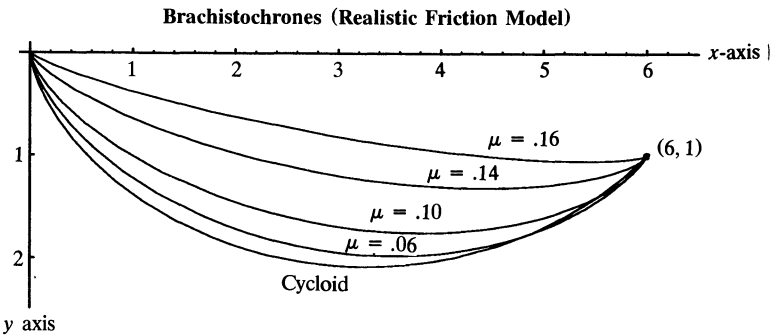


Figure 7. Realistic “frictional” Brachistochrones approaching a straight line

The reader should not be misled into thinking that all of the graphical and physical intuition that has been developed in this article was known to the authors before beginning a computer-aided investigation of the Brachistochrone problem and variations. In fact, this is one of the main points of this article; just as the authors gained valuable insight through computer generated graphics and simulations, so will students at all levels increase their mathematical intuition with these tools and be provided the opportunity to begin exploring questions they would have never even thought of before.

4. ADDITIONAL PROBLEMS. We have looked at other variations of the classic Brachistochrone problem. Many interesting questions can be generated by restricting the class of admissible curves, as with the fastest parabola or n th root problems given earlier. One that we find particularly enjoyable is the “Two Line” Brachistochrone problem: Find the “break” point for the fastest two straight line segment curve joining the origin and the ending point. A similar problem has appeared several times in the problems section of the MAA Monthly; there the question was whether or not the time of travel along the two line segments were equal for the optimal break point (answered in E1255 [1977, 652]). A program like *Mathematica* makes it easy for students to find the break point numerically, which frees them to investigate other questions, such as the problem posed in the Monthly or to look for cases where there is a simple algorithm for finding the break point.

Similar to the constrained variational problem with friction included in the model, is the question of what effect air resistance has on the fastest curve. We are currently attempting to solve this problem numerically so the solution can be

presented graphically in our classes and compared with other curves, although it seems likely that the effect of air resistance may be far less significant than the effect of friction.

The Two Lines problem is included in our Brachistochrone notebook and an article concerning the use of *Mathematica* to investigate more realistic treatments of friction and air resistance is in preparation. We invite questions or discussions on these problems and will send completed or partially completed materials upon request.

REFERENCE

1. N. Ashby, W. E. Brittin, W. F. Love and W. Wyss, Brachistochrone With Coulomb Friction, *American Journal of Physics*, Vol. 43, No. 10 (October 1975), 902–905.

Department of Mathematics & Statistics
California State University, Chico
Chico, CA 95929
lhaws@oavax.csuchico.edu
tkiser@oavax.csuchico.edu

PICTURE PUZZLE

(from the collection of Paul Halmos)



What conspiracy is this?
(see page 344.)

Continued Fractions, Chebychev Polynomials, and Chaos

by William Derrick and Jack Eidswick

1. INTRODUCTION. In this paper we uncover the phenomenon of *chaos* in continued fractions. Our definition of chaos follows that of [4] and our results complement those of [3].

We begin with a calculator/computer investigation of convergence of continued fractions of the following type:

$$\mathcal{F}_2(a) = 2 - \frac{a}{2 - \frac{a}{2 - \frac{a}{\ddots}}} \quad (1)$$

In other words, we calculate the sequence of *partial continued fractions* $t_1 = 2$, $t_2 = 2 - a/2$, $t_3 = 2 - a/(2 - a/2)$, ... for a fixed value of a , and try to make conjectures based on those calculations. For future reference, we note that the sequence $\{t_n\}$ can be expressed iteratively by the difference equation

$$t_{n+1} = 2 - \frac{a}{t_n}, t_1 = 2 \quad (n = 1, 2, \dots). \quad (2)$$

For arbitrary t_1 , (2) is a one-parameter family of discrete dynamical systems, and it is in this context that we will speak of *chaos* in Sections 6 and 7.

If you have a calculator or computer, you can readily check that $\mathcal{F}_2(1) = 1$ and $\mathcal{F}_2(0.98) = 1.14142135624 \dots$. These answers can also be obtained by assuming convergence in (2) to t and solving the quadratic $t^2 = 2t - a$. In general, one can easily prove the following result.

If $a \leq 1$, then the continued fraction (1) converges to $1 + \sqrt{1 - a}$.

But what happens when $a > 1$? Using the quadratic from (2) with $a = 1.02$ yields $\mathcal{F}_2(1.02) = 1 \pm i\sqrt{2}/10$, which is clearly impossible. Thus, these iterations do not converge. Typical t_n for this continued fraction stay close to 1, but, with predictable regularity, gradually decrease until negative, then bounce back to a number greater than 2. A similar phenomenon has been studied for Newton's method in [10].

What are the cluster points of $\{t_n\}$? The graphs in Figure 1 show the sequence $\{t_n\}$ in the viewing window $[0, N] \times [-3.1, 3.2]$ and suggest that there may be many cluster points. On the other hand, graphs away from the line $y = 1$ show very few points, and suggest a sparsity of such points.

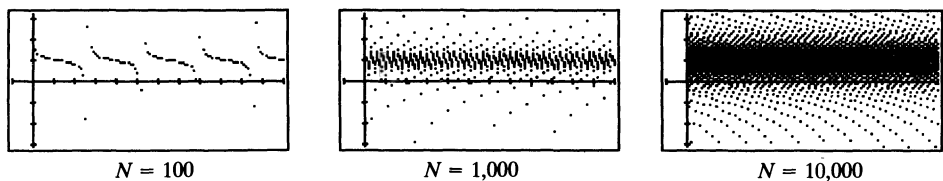


Figure 1

While the case $a = 1.02$ may seem peculiar, investigation at other values of $a > 1$ reveals similar patterns. The graphs in Figure 2 show the sequences $\{t_n\} = \{t_n(a)\}$ in the viewing window $[0, 100] \times [-3.1, 3.2]$ for the indicated values of a .

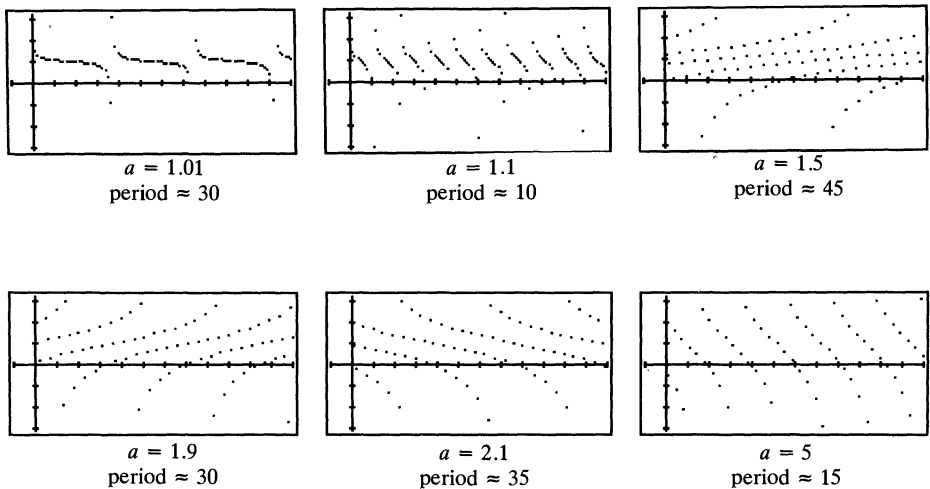


Figure 2

These graphs suggest that there may be periodic values of a in the vicinities of the values shown. Indeed, we will see that this is the case.

Our goals are (i) to determine which values of a lead to periodicity and (ii) to describe the cluster sets for the remaining values of a .

2. CONTINUED FRACTIONS. For a basic treatment of the subject of continued fractions, see [6], [8], or [9]. For a brief history and other information, see [3]. Observe that any continued fraction of the form

$$\begin{aligned}
 x + \frac{y}{x + \frac{y}{x + \frac{y}{x + \dots}}} &= \frac{x}{2} \left[2 - \frac{(-2y/x)}{x + \frac{y}{x + \frac{y}{x + \dots}}} \right] \\
 &= \dots = \frac{x}{2} \left[2 - \frac{a}{2 - \frac{a}{2 - \frac{a}{2 - \dots}}} \right] \quad (3)
 \end{aligned}$$

reduces to a continued fraction of the form (1) with $a = -4y/x^2$. We focus only on continued fractions of the form (1).

3. PERIODIC POINTS OF CONTINUED FRACTIONS. By a *periodic point* of (1), we will mean a point a such that, for some natural number n ,

$$t_{n+k}(a) = t_k(a) \quad \text{for all } k = 1, 2, \dots \quad (4)$$

As above in (2), $t_k(a)$ denotes the k th partial continued fraction of (1). We will allow $t_k(a)$ to assume the value ∞ . For instance, if $a = 2$, then by (2), $t_2(2) = 2$, $t_2(2) = 1$, $t_3(2) = 0$, $t_4(2) = \infty$, and $t_5(2) = 2$; thus, $a = 2$ is a periodic point of period 4.

Theorem 1. *A number a is a periodic point of (1) if and only if $a - 1$ is a zero of the polynomial*

$$P_n(x) = \sum_{k=0}^{\left\lfloor \frac{n-1}{2} \right\rfloor} (-1)^k \binom{n}{2k+1} x^k, \quad (5)$$

where $\lfloor \cdot \rfloor$ denotes the integer part of the number enclosed.

Lemma 1. *If a satisfies (4), then $t_{n-1}(a) = 0$.*

Lemma 2. *If $P_n(x)$ is defined by (5), then*

$$\begin{aligned} P_{n+2}(x) &= 2P_{n+1}(x) - (x+1)P_n(x) \\ P_1(x) &= 1, P_2(x) = 2. \end{aligned} \quad (6)$$

Lemma 3. $t_n(a) = P_{n+1}(a-1)/P_n(a-1)$ for $n = 1, 2, \dots$

Proof of Theorem 1: If a is a periodic point, then a satisfies (4) for some n , and, by Lemma 1, $t_{n-1}(a) = 0$. By back-substituting in (6), we see that not both $P_n(a-1)$ and $P_{n-1}(a-1)$ can be zero. Therefore, $P_n(a-1) = 0$ by Lemma 3. Conversely, if $P_n(a-1) = 0$, then $t_{n-1}(a) = 0$ by Lemma 3, from which it follows that a is a periodic point. ■

Proof of Lemma 1: If (4) holds, $t_{n+1}(a) = t_1(a) = 2$, and, by (2), $t_n(a) = \infty$ and $t_{n-1}(a) = 0$. ■

Proof of Lemma 2: If $P_n(x)$ is defined by (5), then

$$\begin{aligned} Q_n(x) &= P_{n+1}(x) - P_n(x) = \sum_{k=0}^{\lfloor n/2 \rfloor} (-1)^k \binom{n}{2k} x^k \quad \text{and} \\ Q_{n+1}(x) - Q_n(x) &= -x \sum_{k=0}^{\lfloor (n-1)/2 \rfloor} (-1)^k \binom{n}{2k+1} x^k = -xP_n(x). \end{aligned} \quad (7)$$

Therefore, $(P_{n+2}(x) - P_{n+1}(x)) - (P_{n+1}(x) - P_n(x)) = -xP_n(x)$, from which (6) follows. ■

Proof of Lemma 3: By (2), (6), and induction,

$$t_n(a) = 2 - \frac{a}{t_{n-1}(a)} = \frac{2P_n(a-1) - aP_{n-1}(a-1)}{P_n(a-1)} = \frac{P_{n+1}(a-1)}{P_n(a-1)}. \quad \blacksquare$$

The polynomials $P_n(x)$ and $Q_n(x)$ satisfy several interesting identities.

Theorem 2.

- (i) $P_n(x) = P_k(x)P_{n-k+1}(x) - (x+1)P_{k-1}(x)P_{n-k}(x)$.
- (ii) $Q_n(x) = P_k(x)Q_{n-k+1}(x) - (x+1)P_{k-1}(x)Q_{n-k}(x)$.
- (iii) $Q_k^2(x) - xP_k^2(x) = Q_{2k}(x)$.
- (iv) $Q_k^2(x) + xP_k^2(x) = (1+x)^k$.
- (v) $P_{2k}(x) = 2P_k(x)Q_k(x)$.

Proof of Theorem 2:

- (i) For $k = 2$, this is (6). Then by induction,

$$\begin{aligned} P_n &= P_{k-1}P_{n-k+2} - (x+1)P_{k-2}P_{n-k+1} \\ &= P_{k-1}[P_2P_{n-k+1} - (x+1)P_1P_{n-k}] - (x+1)P_{k-2}P_{n-k+1} \\ &= P_{n-k+1}[P_2P_{k-1} - (x+1)P_1P_{k-2}] - (x+1)P_{k-1}P_{n-k}. \end{aligned}$$

- (ii) By (7), $(Q_n - Q_{n-1}) - (Q_{n-1} - Q_{n-2}) = -xQ_{n-2}$, so that

$Q_n = P_2Q_{n-1} - (x+1)P_1Q_{n-2}$. As in (i), the result follows by induction.

- (iii) Let $n = 2k$ in (ii) to get

$$\begin{aligned} Q_{2k} &= P_kQ_{k+1} - (x+1)P_{k-1}Q_k \\ &= P_k(Q_{k+1} - Q_k) + Q_k(2P_k - (x+1)P_{k-1}) - P_kQ_k \end{aligned}$$

from which the identity follows by (6) and (7).

- (iv) From (5) and (7), it follows that $Q_n(x) \pm i\sqrt{x}P_n(x) = (1 \pm i\sqrt{x})^n$.
- (v) Set $n = 2k$ in (i) and use (ii) to obtain

$$\begin{aligned} P_{2k} &= P_k[P_{k+1} - (x+1)P_{k-1}] = 2P_k[2P_k - (x+1)P_{k-1} - P_k] \\ &= 2P_kQ_k. \quad \blacksquare \end{aligned}$$

The identities (iii)–(v) in Theorem 2 bear a surprising resemblance to trigonometric identities if we make the substitution

$$\frac{Q_k(x)}{(1+x)^{k/2}} = \cos k\theta, \quad \frac{\sqrt{x}P_k(x)}{(1+x)^{k/2}} = \sin k\theta.$$

This resemblance is no accident as we shall see in the next section.

4. CHEBYCHEV POLYNOMIALS. The Chebychev polynomials of the first and second kinds are defined, respectively, by

$$\begin{aligned} T_n(\cos \theta) &= \cos n\theta \quad \text{and} \\ U_n(\cos \theta) &= \frac{\sin(n+1)\theta}{\sin \theta} \quad \text{for } n = 1, 2, \dots \end{aligned} \quad (8)$$

From these, one may derive the following explicit representations:

$$\begin{aligned} T_n(x) &= \sum_{m=0}^{[n/2]} \binom{n}{2m} x^{n-2m} (x^2 - 1)^m \\ U_n(x) &= \sum_{m=0}^{[n/2]} \binom{n+1}{2m+1} x^{n-2m} (x^2 - 1)^m. \end{aligned} \quad (9)$$

Chebyshev polynomials, along with Jacobi polynomials, Gegenbauer polynomials, Hermite polynomials, Laguerre polynomials, and Legendre polynomials, share the distinction of being known as “classical polynomials”. A great deal is known about these polynomials; see, e.g., [7, p. 207, 257], and [5]. For our purposes, we only need the above representations.

Theorem 3. *The set of periodic points of (1) is equal to*

$$\mathcal{P} = \left\{ 1 + \tan^2 \frac{k\pi}{n+1} : k = 1, 2, \dots, \left[\frac{n}{2} \right]; n = 1, 2, 3, \dots \right\}.$$

Corollary 1. *The periodic points of (1) are dense in the interval $(1, \infty)$.*

Proof of Theorem 3: By (5) and (9), $x^n P_{n+1}((1 - x^2)/x^2) = U_n(x)$, and, therefore, by (8),

$$\cos^n \theta \cdot P_{n+1}(\tan^2 \theta) = \frac{\sin(n+1)\theta}{\sin \theta}, \quad (10)$$

from which it follows that the zeros of $P_{n+1}(x)$ are $x = \tan^2(k\pi/(n+1))$, $k = 1, 2, \dots, [n/2]$. ■

Proof of Corollary 1: Since the set

$$\{k\pi/(n+1) : k = 1, 2, \dots, [n/2]; n = 1, 2, \dots\}$$

is dense in $(0, \pi/2)$ and the function $1 + \tan^2 x$ is continuous in $(0, \pi/2)$, the periodic points are dense in $(1, \infty)$. ■

5. NONPERIODIC POINTS OF CONTINUED FRACTIONS. We will now determine the cluster sets of $\{t_n(a)\}$ for nonperiodic values of a .

Theorem 4. *If $a > 1$ is a nonperiodic point of (1), then the sequence $\{t_n(a)\}$ of partial continued fractions clusters at every real number.*

Lemma 4. *If $0 < \alpha < \pi/2$ and $\alpha \notin \{k\pi/(n+1) : k = 1, 2, \dots, [n/2]; n = 1, 2, \dots\}$, then the set $\{\tan(n\alpha) : n = 1, 2, \dots\}$ is dense in $(-\infty, \infty)$.*

Proof of Lemma 4: The lemma is clearly equivalent to the following statement:

If $e^{i n \alpha} \neq 1$ for all $n = 1, 2, \dots$, then $S = \{e^{i n \alpha} : n = 1, 2, \dots\}$ is dense in the unit circle T .

To prove this statement, suppose $T - S$ contains an arc $\mathcal{A} = \{e^{it} : t_1 < t < t_2\}$ and consider arcs $\mathcal{A}_+ = \{e^{it} : 0 < t < t_2 - t_1\}$ and $\mathcal{A}_- = \{e^{-it} : 0 < t < t_2 - t_1\}$. Since arcs joining successive powers of the $e^{i\alpha}$ all have length equal to α , $t_2 - t_1$ is

necessarily less than α . The arc A_+ can contain no point of the form $e^{i n \alpha}$ because, if it did, we could write $e^{i n \alpha} = e^{i \beta}$, where $\beta < t_2 - t_1$, from which we would get that some power of $e^{i \beta}$ would be in A ; i.e., $(e^{i \beta})^k = e^{i k n \alpha}$ would be in A for some k , contrary to our assumption. Similarly, A_- can contain no point of the form $e^{i n \alpha}$. In other words, $T - S$ contains $A_+ \cup A_-$. If, in addition, we have $e^{i n \alpha} \neq 1$ for $n = 1, 2, \dots$, then $T - S$ contains $\{e^{i t} : |t| < t_2 - t_1\}$, an arc having twice the length of A . Repeating this process gives the existence of an arc in $T - S$ which has length greater than α , contrary to the above observation. The lemma follows. ■

Proof of Theorem 4: By Lemma 3 and (10),

$$t_n(a) = t_n(\tan^2 \alpha + 1) = \frac{P_{n+1}(\tan^2 \alpha)}{P_n(\tan^2 \alpha)} = \frac{\sin(n + 1) \alpha}{\sin(n \alpha) \cos \alpha} = 1 + \frac{\tan \alpha}{\tan(n \alpha)},$$

where $\alpha = \tan^{-1} \sqrt{a - 1} > 0$ and

$$\alpha \notin \left\{ \frac{k \pi}{n + 1} : k = 1, 2, \dots, \left[\frac{n}{2} \right]; n = 1, 2, \dots \right\}.$$

Therefore, by Lemma 4, the points $t_n(a)$ are dense in $(-\infty, \infty)$. ■

6. DYNAMICAL SYSTEMS TERMINOLOGY. If f is a continuous function on a metric space X , then by a *discrete dynamical system* is meant the family of sequences

$$x, f(x), f(f(x)), \dots, f^{(n)}(x), \dots$$

where $x \in X$. A point x is called a *periodic point* of the system if $f^{(n)}(x) = x$ for some natural number n .

After noting the nonexistence of a uniformly accepted definition of *chaos* and quoting Bill Thurston: “to call your field ‘chaotic’ is an admission of defeat from the outset”, Devaney [4, p. 17] adopts the following three conditions for chaos:

1. f has sensitive dependence on initial conditions,
2. f is topologically transitive,
3. the periodic points are dense in X .

Condition 1 means that there exists a positive number δ such that, for any $x \in X$ and any neighborhood U of x , there exists $y \in U$ and a natural number n such that $\text{dist}(f^{(n)}(x) - f^{(n)}(y)) > \delta$. Condition 2 means that for any pair of open sets $U, V \subseteq X$, there exists n such that $f^{(n)}(U) \cap V \neq \emptyset$. Conditions 1 and 2 are the chaotic portions of Devaney’s definition. Roughly speaking, Condition 1 says that if you start with any orbit, then there are other orbits that start arbitrarily close to it, but eventually stray from it by at least δ units; and Condition 2 says that if you start with any two locations, then there is an orbit leading from one to the other. Condition 3 establishes an element of regularity in the definition. It is known that Conditions 2 and 3 imply Condition 1 (see [2]).

7. CHAOS AND CONTINUED FRACTIONS. The results of Sections 4 and 5 are quite striking and suggest the presence of an underlying chaotic dynamical system. Below, we identify a one-parameter family of discrete dynamical systems which collectively set a framework for our results. In Theorem 5, we give an analysis of these systems according to whether $a \in \mathcal{P}$ or $a \notin \mathcal{P}$, where \mathcal{P} is the set

described in Theorem 3. The results of Theorem 5, along with the above discussion, justify our contention that the behavior of continued fractions of the form (1) is chaotic. In part (2) of Theorem 5, we actually show more than what is needed to verify the condition of sensitive dependence. What we show is that, in a sense, the iterates separate exponentially.

For $a > 1$, define $T(x) = 2 - a/x$ for $x \in \mathbb{R}, x \neq 0$, and consider the following dynamical system:

$$x, T(x), T(T(x)), \dots, T^{(n)}(x), \dots \tag{11}$$

Theorem 5. (1) If $a \in \mathcal{P}$, then (i) T does not have sensitive dependence on initial conditions, (ii) T is not topologically transitive, and (iii) every real number $x \neq 0$ is a periodic point.

(2) If $a \notin \mathcal{P}$, then the situation is reversed: (i) T has sensitive dependence on initial conditions, (ii) T is topologically transitive, and (iii) no real number $x \neq 0$ is a periodic point.

The following lemma is an easy generalization of Lemma 3.

Lemma 5. $T^{(n)}(x) = [xP_{n+1}(a - 1) - aP_n(a - 1)]/[P_{n+1}(a - 1) + (x - 2)P_n(a - 1)]$ for $n = 1, 2, 3, \dots$

Proof of Theorem 5, part (1): Assume $a \in \mathcal{P}$ with period n , so that $P_n(a - 1) = 0$ and $P_k(a - 1) \neq 0$ for $k = 1, 2, \dots, n - 1$. Define $\lambda_k = P_{k+1}(a - 1)/P_k(a - 1)$ for $k = 1, 2, \dots, n - 1$. Then, by Lemma 5,

$$T^{(k)}(x) = \frac{x\lambda_k - a}{\lambda_k + (x - 2)} \quad \text{for } k = 1, 2, \dots, n - 1, \text{ and}$$

$$T^{(n)}(x) = x,$$

from which (iii) follows.

Also, for $\delta > 0$, choose $x \neq 2 - \lambda_k$ for $k = 1, 2, \dots, n - 1$, and $\varepsilon > 0$ so small that $|T^{(k)}(x) - T^{(k)}(y)| < \delta$ whenever $|x - y| < \varepsilon$ and $k = 1, 2, \dots, n$ (which is possible by continuity of $T^{(k)}$ at x). Condition (i) follows by periodicity.

Let $\delta = 1$. Then for x and ε as above, $|x - y| < \varepsilon$, and $m = 1, 2, \dots$, the sequence $\{T^{(m)}(y)\}$ is bounded by $\text{Max}\{|T^{(k)}(x)| + 1: k = 1, 2, \dots, n\}$; hence, (ii) follows. ■

Lemma 6. If $a \notin \mathcal{P}$, there exists a sequence $\{\lambda_n\}$, dense in \mathbb{R} , such that

$$T^{(n)}(x) = \frac{x\lambda_n - a}{\lambda_n + x - 2}.$$

Proof of Lemma 6: Define $\alpha = \tan^{-1}\sqrt{a - 1}$. By Lemma 5 and the proof of Theorem 4, $T^{(n)}(x)$ has the desired form with

$$\lambda_n = \frac{P_{n+1}(\tan^2 \alpha)}{P_n(\tan^2 \alpha)} = 1 + \frac{\tan \alpha}{\tan(n\alpha)}.$$

The result then follows from Lemma 4. ■

Proof of Theorem 5, part (2): Assume $a \notin \mathcal{P}$. If $|x| > 3a$, then $|Tx| < 3 < 3a$, so every sequence $\{T^{(k)}x\}$ intersects the interval $J = [-3a, 3a]$. If we can show

sequences beginning in J have sensitive dependence on initial conditions, then all sequences will have sensitive dependence as they intersect J . Let $0 < \varepsilon < \min(1, (a - 1)/5)$. Since $\{\lambda_n\}$ is dense in \mathbb{R} , there exists $N = N(\varepsilon)$ such that for every x in J , we can find λ_n such that $\varepsilon/4 < |x - (2 - \lambda_n)| < \varepsilon/2$ and $n < N$. Select any y such that $\varepsilon/16 < |x - y| < \varepsilon/8$, then

$$\begin{aligned} |T^{(n)}(x) - T^{(n)}(y)| &= \frac{|x - y|[(\lambda_n - 1)^2 + (a - 1)]}{|x - (2 - \lambda_n)||y - (2 - \lambda_n)|} > \frac{(\varepsilon/16)(a - 1)}{(\varepsilon/2)(5\varepsilon/8)} \\ &= \frac{a - 1}{5\varepsilon} > 1 > 8\frac{\varepsilon}{8} > e^{\ln 8}|x - y| > e^{(1/N \ln 8)n}|x - y| \\ &= e^{\alpha n}|x - y|, \end{aligned}$$

where $\alpha = (\ln 8)/N > 0$. This gives exponential divergence since the Lyapunov exponent α is positive. See [1, p. 85].

Also, from Lemma 6 and the fact that $x(2 - x) - a \neq 0$, we see that $\{T^{(n)}(x)\}$ is dense in \mathbb{R} for any x and, therefore, (ii) and (iii) follow. ■

Remark. If a is complex and $\text{Im } a \neq 0$, it is not hard to show that the continued fraction (1) converges. Thus, chaos occurs only when $a > 1$.

REFERENCES

1. Baker, G. L. and Gollub, J. P., *Chaotic Dynamics an Introduction*, Cambridge University Press, Cambridge, 1990.
2. Banks, J., Brooks, J., Cairns, G., and Stacey, P., On Devaney's Definition of Chaos, *Amer. Math. Monthly*, 99 (1992) 332–334.
3. Corless, R. M., Continued Fractions and Chaos, *Amer. Math. Monthly* 99 (1992) 203–215.
4. Devaney, Robert L., *Chaotic Dynamical Systems*, Second Edition, Addison-Wesley, Redwood City, California, 1991.
5. Fox, L. and Parker, I. B., *Chebyshev Polynomials in Numerical Analysis*, Oxford University Press, London, 1968.
6. Hardy, G. H. and Wright, E. M., *Introduction to the Theory of Numbers*, Oxford University Press, London, 1945.
7. Magnus, W., Oberhettinger, F., and Soni, R. P., *Formulas and Theorems for the Special Functions of Mathematical Physics*, Springer-Verlag, Vol. 52, New York, 1966.
8. Niven, Ivan and Zuckerman, Herbert, *An Introduction to the Theory of Numbers*, John Wiley & Sons, New York, 1960.
9. Olds, C. D., *Continued Fractions*, Random House, New York, 1963.
10. Strang, G., A Chaotic Search for i , *College Math. Journal*, 22(1991) 3–12.

Department of Mathematical Sciences
The University of Montana
Missoula, MT 59812-1032
derrick@selway.umt.edu
ma_jae@selway.umt.edu

Answer to Picture Puzzle (p. 336)

R. D. Anderson, R. H. Bing, and R. S. Palais
at a meeting of the Council of the AMS (in
Toronto, August 1967).

A Relation Between Partitions and the Number of Divisors

Wang Zheng Bing, Robert Fokkink and Wan Fokkink

A sum of positive natural numbers adding up to n is called a *partition* of n . For instance, $1 + 2 + 4$ is a partition of 7. As none of the summands 1, 2, 4 are equal, this is called a partition *into unequal parts*. There are five partitions of 7 into unequal parts:

$$1 + 2 + 4, \quad 1 + 6, \quad 2 + 5, \quad 3 + 4, \quad 7.$$

Since the partitions $1 + 2 + 4$ and 7 contain an odd number of summands, they are called *odd* partitions, whereas the other three partitions are called *even*. Add the smallest numbers of the odd partitions, $1 + 7 = 8$, and do the same for the smallest numbers of the even partitions, $1 + 2 + 3 = 6$. The difference between these two sums, $8 - 6 = 2$, is exactly the number of divisors of the prime 7.

In the sequel, $p(n)$ denotes the sum of the smallest numbers of odd partitions of n minus the smallest numbers of even partitions of n , and $d(n)$ denotes the number of divisors of n . For small numbers n , it is easy to check that $p(n)$ equals $d(n)$. This is not a coincidence; we shall see that it is a general relation between the smallest numbers of partitions into unequal parts and the number of divisors.

Theorem. $p(n) = d(n)$ for all positive natural numbers n .

In order to prove this theorem, we introduce the sum of polynomial quotients

$$P_n(X) = \sum_{i=0}^{n-1} \frac{(1 - X^{i+1})(1 - X^{i+2}) \cdots (1 - X^n)}{1 - X^{n-i}}$$

for positive natural numbers n . At each consecutive quotient, the degree of the denominator decreases by one and the leftmost factor in the numerator drops out. Fix an $m = 1, \dots, n$. We shall show that the coefficient α_m for X^m in $P_n(X)$ equals $d(m) - p(m)$.

First, we determine the contributions from the separate quotients of $P_n(X)$ to α_m . Fix an $i = 0, \dots, n - 1$, and replace the denominator $1/(1 - X^{n-i})$ in the i th quotient of $P_n(X)$ by its power series (which converges for $|X| < 1$). Hence, the i th quotient of $P_n(X)$ takes the form

$$(1 - X^{i+1}) \cdots (1 - X^n)(1 + X^{n-i} + X^{2(n-i)} + \cdots).$$

Since $m \leq n$, the contributions from this product to α_m stem either from $(1 - X^{i+1}) \cdots (1 - X^n)$ or from $(1 + X^{n-i} + X^{2(n-i)} + \cdots)$. Now, we collect the contributions to α_m of these two types of terms.

- Clearly, the series $(1 + X^{n-i} + X^{2(n-i)} + \cdots)$ contributes $+1$ to the coefficient α_m of X^m if and only if $n - i$ is a divisor of m .
As i increases from 0 to $n - 1$, the number $n - i$ decreases from n to 1 . In this range there are $d(m)$ numbers which divide m , so there are $d(m)$ series $(1 + X^{n-i} + X^{2(n-i)} + \cdots)$ for $i = 0, \dots, n - 1$ which contribute $+1$ to α_m . These contributions together sum up to $d(m)$.
- If we decompose the product $(1 - X^{i+1}) \cdots (1 - X^n)$, this results into terms $(-1)^l X^{k_1 + \cdots + k_l}$ for all sequences of numbers $i + 1 \leq k_1 < \cdots < k_l \leq n$. So this product contributes $+1$ to α_m for each even partition of m with terms greater than i , and it contributes -1 to α_m for each odd partition of m with terms greater than i .
So for each even partition of m with smallest term k , the products $(1 - X^{i+1}) \cdots (1 - X^n)$ for $i = 0, \dots, k - 1$ contribute $+1$ to α_m . These contributions together sum up to k . Even so, for each odd partition of m with smallest term k , the products $(1 - X^{i+1}) \cdots (1 - X^n)$ for $i = 0, \dots, k - 1$ contribute -1 to α_m . These contributions together sum up to $-k$. So in total, these contributions to α_m sum up to $-p(m)$.

Hence, we have found that α_m equals $d(m) - p(m)$ for $m = 1, \dots, n$. So to prove our main theorem, it suffices to prove the following proposition.

Proposition. $P_n(X)$ equals n for all $n \geq 1$.

Proof: Note that $P_1(X) = 1$. To prove the proposition, we show that the difference between $P_{n+1}(X)$ and $P_n(X)$ is equal to 1 .

Shifting the index i of the sum $P_n(X)$ by one, $P_{n+1}(X) - P_n(X)$ takes the form

$$\sum_{i=0}^n \frac{(1 - X^{i+1}) \cdots (1 - X^{n+1})}{1 - X^{n+1-i}} - \sum_{i=1}^n \frac{(1 - X^i) \cdots (1 - X^n)}{1 - X^{n+1-i}}.$$

In the second sum, we can also start the index i at 0 , because its quotient for $i = 0$ equals zero. Now, collecting quotients of equal denominator gives

$$\sum_{i=0}^n (X^i - X^{n+1}) \frac{(1 - X^{i+1}) \cdots (1 - X^n)}{1 - X^{n+1-i}}.$$

The denominator $1 - X^{n+1-i}$ divides the factor $X^i - X^{n+1}$ in the numerator, so this sum equals

$$\sum_{i=0}^n X^i (1 - X^{i+1}) \cdots (1 - X^n).$$

Denote this polynomial by $Q_n(X)$. The result will follow if $Q_n(x) = 1$ for all n . Again, we use induction. $Q_1(x) = (1 - X) + X = 1$, and isolating the term with $i = n + 1$ in the expression for $Q_{n+1}(x)$ yields the relation $Q_{n+1}(x) = (1 - X^{n+1})Q_n(x) + X^{n+1}$, which provides the inductive step. ■

ACKNOWLEDGMENTS. Jos van Wamel is thanked for helpful comments, and special thanks go to Joost Peeters, who encountered the polynomial $P_n(X)$ in an optimal control model [1].

1. J. Peeters and O. Ciftcioglu. Statistics on exponential averaging of periodograms, to appear in *IEEE J. Sign. Proc.*

Wang Z. B.:
Department of Civil Engineering
Delft University of Technology
P.O. Box 5048
2600 GA Delft
The Netherlands

R. Fokkink:
Department of Mathematics
Delft University of Technology
P.O. Box 3051
2600 GA Delft
The Netherlands

W. Fokkink:
Department of Computer Science
Centre for Mathematics and Computer Science (CWI)
P.O. Box 94079
1090 GB Amsterdam
The Netherlands

Answers to Two Questions Concerning Quotients of Primes

Paolo Starni

We will consider some open questions concerning quotients of primes posed in [1] by D. Hobby and D. M. Silberger.

Corollary 3 of this note will solve Open Problem 2 of their paper, and our Theorem 1 represents some progress on their Open Problem 1. \mathbb{R}^+ and \mathbb{N} denote, as usual, the set of all positive real numbers and the set of all positive integers (0 excluded). If $S \subseteq \mathbb{N}$, we indicate with $\mathbf{F}(S)$ the set of all quotients p/q for which $\{p, q\} \subseteq S$ and $p \neq q$; so, for instance, $\mathbf{F}(\mathbb{N}) = \mathbb{Q}^+ - \{1\}$, where \mathbb{Q}^+ represents the set of all positive rational numbers (0 excluded).

Open Problem One is:

Characterize the family of all $S \subseteq \mathbb{N}$, S infinite, for which $\mathbf{F}(S)$ is dense in \mathbb{R}^+ .

We note that, by the definition of $\mathbf{F}(S)$, density in $]0, 1]$ implies density in \mathbb{R}^+ . It is convenient to assume here the following definition of density (see [2] *passim*): Let $X \subseteq Y \subseteq \mathbb{R}^+$. The set X is dense in Y if and only if for every $y \in Y$ there exists a sequence x_n in X such that $\lim_{n \rightarrow +\infty} x_n = y$.

The following theorem gives a sufficient condition for $\mathbf{F}(S)$ to be dense.

Theorem 1. *If there exists a strictly increasing sequence in S , p_n , such that $\lim_{n \rightarrow +\infty} p_{n-1}/p_n = 1$, then $\mathbf{F}(S)$ is dense in \mathbb{R}^+ .*

Proof: Let $x \in]0, 1]$ be given, and choose k in such a way that $x p_k > p_1$. For any $n > k$, we may choose m so that $p_{m-1} < x p_n \leq p_m$, making m a function of n . We then have

$$0 \leq (p_m - x p_n)/p_n < (p_m - p_{m-1})/p_n \leq (p_m - p_{m-1})/p_m = 1 - p_{m-1}/p_m.$$

When n goes to positive infinity, m goes to positive infinity as well, so we have

$$0 \leq \lim_{n \rightarrow +\infty} (p_m - x p_n)/p_n \leq 1 - \lim_{m \rightarrow +\infty} (p_{m-1}/p_m) = 1 - 1 = 0.$$

Thus

$$\lim_{n \rightarrow +\infty} ((p_m/p_n) - x) = 0 \text{ and so } \lim_{n \rightarrow +\infty} p_m/p_n = x. \quad \blacksquare$$

From Theorem 1 we deduce the density of $\mathbf{F}(S)$ in \mathbb{R}^+ if $S = \{a + nb; a, b \in \mathbb{N}, n = 1, 2, \dots\}$. Since the set of all positive integers odd (\mathbb{O}) or even (\mathbb{E}) may be considered as an arithmetic progression, we obtain that $\mathbf{F}(\mathbb{O})$ and $\mathbf{F}(\mathbb{E})$ are both dense in \mathbb{R}^+ . So also is $\mathbf{F}(\mathbb{N})$, but in this case the result is trivial since it is well known that \mathbb{Q}^+ (or $\mathbb{Q}^+ - \{1\}$) is dense in \mathbb{R}^+ . Besides, we obtain that:

Corollary 2. $\mathbf{F}(\mathbb{P})$ is dense in \mathbb{R}^+ , where \mathbb{P} denotes the set of all primes.

(This corollary represents another proof of Theorem 4 in [1].)

Proof: Let p_n be the sequence of all primes in increasing order. The Prime Number Theorem (cited also in [1]) implies that asymptotically $p_n = n \log n$ (for the proof see [3], p. 10) and so $\lim_{n \rightarrow +\infty} p_{n-1}/p_n = 1$. The proof now follows from Theorem 1. \blacksquare

$\mathbf{D}(a, b)$ denotes the set of all primes which belong to the arithmetic progression $\{a + nb; a, b \in \mathbb{N}, n = 1, 2, \dots\}$. If a and b are coprime, then $\mathbf{D}(a, b)$ is infinite (this is Dirichlet's theorem cited also in [1]).

Corollary 3. $\mathbf{F}(\mathbf{D}(a, b))$ is dense in \mathbb{R}^+ , whenever a and b are coprime.

(This statement answers affirmatively the question posed in [1] as Open Problem Two.)

Proof: Let $\pi(x)$ be the number of primes $\leq x$. When a and b are coprime we indicate by $\pi_{a,b}(x)$ the number of primes in the set $\{p_n; n \in \mathbb{N}\} = \mathbf{D}(a, b)$ that do not exceed x . We have that $\pi_{a,b}(x)$ is asymptotically equal to $\pi(x)/\phi(b)$, where ϕ is Euler's totient function (see [4], p. 214). Also in this case by the Prime Number Theorem one obtains that asymptotically $p_n = n \log n$ and the proof follows as in Corollary 2. \blacksquare

Theorem 4. The converse of Theorem 1 is false.

Proof: Let $S = \bigcup_{n=1}^{\infty} S_n$, where $S_n = \{2^{2n}, 2^{2n} + 1, \dots, 2^{2n+1} - 1\}$. $\mathbf{F}(S)$ is dense in \mathbb{R}^+ , but $\lim_{n \rightarrow +\infty} (p_{n-1}/p_n) \neq 1$ for any strictly increasing sequence p_n from S .

1) To see that $\mathbf{F}(S)$ is dense let $x \sim S$ mean that x can be approximated arbitrarily closely by p_n/q_n with $p_n, q_n \in S$ and $p_n \neq q_n$. We have the following points.

(i) If $x \sim S$, then $4x, x/4 \sim S$.

(ii) If $x \in [1/4, 1/2]$ or $x \in [1/2, 1]$, then $x \sim S$. We use the dyadic expansion $x = \sum_{j=1}^{\infty} (a_j/2^j)$, $a_j \in \{0, 1\}$.

If $x \in [1/4, 1/2]$, then $x = \frac{1}{4} + \sum_{j=3}^{\infty} (a_j/2^j)$; we must show that $x_n = \frac{1}{4} + \sum_{j=3}^{2n} (a_j/2^j)$ belongs, for $n \geq 2$, to $\mathbf{F}(S)$. We obtain

$$x_n = \frac{2^{2(n-1)} + (a_3 2^{2n-3} + \cdots + a_{2n})}{2^{2n}} = \frac{2^{2(n-1)} + s}{2^{2n}}.$$

Since $0 \leq s \leq 1 + 2 + \cdots + 2^{2n-3} = 2^{2n-2} - 1 < 2^{2n-2}$, $x_n \in \mathbf{F}(S)$.

If $x \in [1/2, 1]$, then $x = \frac{1}{2} + \sum_{j=2}^{\infty} (a_j/2^j)$; when $n \geq 1$

$$x_n = \frac{1}{2} + \sum_{j=2}^{2n+1} \frac{a_j}{2^j} = \frac{2^{2n} + a_2 2^{2n-1} + \cdots + a_{2n+1}}{2^{2n+1}} = \frac{2^{2n} + s_n}{2^{2n+1}}$$

where $0 \leq s_n \leq 2^{2n-1} + \cdots + 1 = 2^{2n} - 1 < 2^{2n}$; thus

$$\lim_{n \rightarrow +\infty} y_n = \frac{2^{2n} + s_n}{2^{2n+1} - 1} = x_n \frac{2^{2n+1}}{2^{2n+1} - 1} = x \quad \text{and} \quad y_n \in \mathbf{F}(S)$$

(iii) if $x \in]0, 1[$, then $x \sim S$.

If $x \in [1/2^{2p}, 1/2^{2p-1}]$ for $p \geq 2$, then $2^{2(p-1)}x \in]1/4, 1/2]$ and $x \sim S$ by (1).

If $x \in [1/2^{2p+1}, 1/2^{2p}]$ for $p \geq 2$, then $2^{2p}x \in]1/2, 1]$ and $x \sim S$ by (1).

As we noted at the start of this proof, (iii) implies that $\mathbf{F}(S)$ is dense.

2) For the other half of the proof, note that if $p, q \in S$, $p < q$, then, setting $p = 2^{2k} + s$ and $q = 2^{2h} + t$, we get $k \leq h$. If $k < h$, then

$$p/q \leq \frac{2^{2(h-1)} + s}{2^{2h} + t} = \frac{1/4 + s 2^{-2h}}{1 + t 2^{-2h}} < \frac{1}{4} + 2^{2(k-h)} \leq \frac{1}{2}.$$

Now let p_n be a strictly increasing sequence in S .

We assume that $\lim_{n \rightarrow +\infty} (p_{n-1}/p_n) = 1$; then it must be that for n sufficiently large, $\frac{1}{2} < (p_{n-1}/p_n) < 1$. So $k = h$ is constant and that is impossible. ■

ACKNOWLEDGMENT. I thank Prof. Piero Plazzi (University of Bologna) and the referee for useful comments and advice.

REFERENCES

1. D. Hobby, D. M. Silberger, Quotients of Primes, *Amer. Math. Monthly*, vol. 100, n. 1, Jan. 1993, p. 50.
2. T. A. Apostol, *Mathematical Analysis*, second edition, Addison-Wesley, Reading, 1977.
3. G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers*, fifth edition, Clarendon Press, Oxford, 1979.
4. P. Ribenboim, *The Book of Prime Number Records*, second edition, Springer-Verlag, New York, 1989.

Liceo Virgilio
Piazza Città, 5
39049 Vipiteno
Bolzano
Italy

Avoiding the Exchange Lemma

James Ford

In finite-dimensional vector space theory, before defining dimension (as the size of a basis) it is necessary to show that all bases have the same number of members. This is usually done by appealing either to the exchange lemma or to the fact that a system of more than n homogeneous linear equations in n unknowns has a nontrivial solution. When the underlying field has zero characteristic, however, the following more direct approach is available.

Let V be a vector space over a field of characteristic zero. A set $\{u_1, u_2, \dots, u_n\}$ of vectors is a **basis** for V if each $v \in V$ can be written uniquely in the form $v = \sum_{i=1}^n \lambda_i u_i$, where the λ_i are scalars.

Lemma. *All bases for V have the same number of members.*

Proof: Let $\{u_1, u_2, \dots, u_n\}$ and $\{v_1, v_2, \dots, v_m\}$ be bases for V . Then there exist unique scalars a_{ij}, b_{ij} ($1 \leq i \leq n, 1 \leq j \leq m$) such that

$$u_i = \sum_{j=1}^m a_{ij} v_j, \quad v_j = \sum_{i=1}^n b_{ji} u_i$$

for each i and j . It follows that for $1 \leq i \leq n$

$$u_i = \sum_{j=1}^m a_{ij} v_j = \sum_{j=1}^m a_{ij} \sum_{k=1}^n b_{jk} u_k$$

i.e.

$$u_i = \sum_{k=1}^n \sum_{j=1}^m a_{ij} b_{jk} u_k.$$

Since $\{u_1, u_2, \dots, u_n\}$ is a basis, the two expressions above for u_i must be the same. Equate the coefficients of u_i on both sides:

$$1 = \sum_{j=1}^m a_{ij} b_{ji} \quad (1 \leq i \leq n)$$

and sum over i :

$$n = \sum_{i=1}^n \sum_{j=1}^m a_{ij} b_{ji}.$$

Similarly,

$$m = \sum_{j=1}^m \sum_{i=1}^n b_{ji} a_{ij}.$$

Since the two double sums are rearrangements of each other, it follows that $m = n$.

Remarks

1. The proof consists of the observation that $\text{Tr}(AB) = \text{Tr}(BA)$, where $A = [a_{ij}]$ and $B = [b_{ji}]$. Note that if the field had nonzero characteristic then the argument would merely establish that the sizes of the two bases differed by a multiple of the characteristic.
2. The same argument can be used in an obvious way to define the degree of a finite extension of a field of characteristic zero. This leads directly to a proof of the double extension theorem: if E is a field of characteristic zero, F is a finite extension of E and G is a finite extension of F , then G is a finite extension of E and $[G : E] = [G : F][F : E]$. This approach can be used to demonstrate the impossibility of the duplication of the cube and the trisection of the angle using straightedge and compass to students who have no knowledge of linear algebra.

*Department of Mathematics and Statistics
University of Newcastle upon Tyne
Newcastle NE1 7RU
United Kingdom
jim.ford@newcastle.ac.uk*

Intervals Contained in Arithmetic Combinations of Sets

Stephen Silverman

If A and B are subsets of \mathbf{R} , then the set $A + B$ is the set of all sums $a + b$ for $a \in A$ and $b \in B$. The sets $A - B$, $A \cdot B$, and A/B are defined similarly (with no dividing by zero).

Theorem (Steinhaus). *If A and B have positive measure then $A + B$ and $A - B$ each contain an interval of positive length.*

See, for example, [2, p. 143]. This result of Steinhaus has been generalized, by Weil, to locally compact groups ([1, p. 296]). For the topological group $(\mathbf{R} \setminus \{0\}, \cdot)$ the invariant (Haar) measure of a set A is given by $\int_A |x|^{-1} dx$, and thus this measure has the same sets of positive measure as does Lebesgue measure. Since an open set in $\mathbf{R} \setminus \{0\}$ is open in \mathbf{R} , Steinhaus's theorem could be augmented to include the sets $A \cdot B$ and A/B in its conclusion as well.

The conclusion doesn't necessarily hold if the sets have measure zero, even if they are closed and uncountable. The existence of closed, uncountable, and independent sets is shown in [4, p. 103], where A is independent if whenever x_i are distinct elements of A and n_i are in \mathbf{Z} , then $n_1 x_1 + \cdots + n_k x_k = 0$ implies that each n_i is 0. So if A is independent and $A + A$ contains an interval and c is a nonzero member of A , then there is a rational p/q with p odd and q even so that $(p/q)c$ is in $A + A$. This means there are a and b in A with $a + b = (p/q)c$ or $qa + qb - pc = 0$ which contradicts independence.

However in the case that A and B are the Cantor set C ([2, p. 71]) we have $C + C = [0, 2]$ and $C - C = [-1, 1]$. This can be seen by adding and subtracting members of C in base 3 recalling that C is the set of all reals in $[0, 1]$ with only 0's and 2's in their ternary representations.

A fascinating observation by Ray Mayer [3] is that $C \cdot C$, though not all of $[0, 1]$ (e.g., $2/5$ is missing), does contain intervals, has measure about .80955, and has countable boundary. The first two facts are not too difficult and make nice Cantor set exercises.

The purpose of this note is to point out that there is a well known class of sets (whether measure zero or not) that have the maximum possible sum, difference, product, and quotient. Our result, which is a further extension of Steinhaus's theorem and which seems to have gone unnoticed, concerns the "arithmetic" of dense G_δ sets.

Recall that A has measure 0 if for each $n > 0$ we can find an open set U_n containing A with total length less than $1/n$. The intersection of the U_n is then a G_δ (the countable intersection of open sets) of measure 0. If A also contains all rational numbers, then it is a dense G_δ of measure 0.

Theorem. *Let G and H be dense G_δ sets in non-empty open intervals I and J respectively. If $\&$ is any one of the four arithmetic operations $+$, $-$, \cdot or $/$, then*

$$G \& H = I \& J.$$

except that in the case of multiplication and division 0 might be in $I \& J$ but not in $G \& H$.

Proof: We consider the simplest case first: $I = J = \mathbf{R}$ and $\& = +$. Let r be in \mathbf{R} and $f(x) = r - x$. Since f is a homeomorphism, $f(H)$ is a dense G_δ set. By the Baire Category Theorem [2, p. 68] $f(H)$ is of the 2nd category and thus cannot lie in $\mathbf{R} \setminus G$ which is of first category. This implies that there is an $x \in H$ such that $f(x)$ is in G , but $f(x) + x = r$ so $r \in G$ finishing this case.

For arbitrary I and J let r be in $I + J$. Then the set $X = I \cap (r - J)$ is a non-empty open interval with $f(H) \cap X$ and $G \cap X$ dense G_δ sets in X , and we can proceed as above. When the operation is subtraction let $f(x) = x + r$, for multiplication let $f(x) = r/x$, and for division let $f(x) = rx$, with $r \neq 0$ in the latter two cases, insuring that f is still a homeomorphism. ■

One might wonder whether the hypothesis of the theorem can be weakened. Consider the following example, which is an application of the above Theorem.

Example. There exists a set A of 2nd category such that $A + A$ contains no interval.

In fact we will show the existence (using the axiom of choice) of an independent set of 2nd category.

First, we observe that the cardinality of any dense G_δ , G is c . If $\text{card}(G) = z$, then $z \leq \text{card}(G + G) \leq \text{card}(G \times G) = z \cdot z = z$, but by the above theorem $G + G$ contains an interval so $z = c$. Second, we note that there are exactly c dense G_δ 's. Since there are c open sets and each G_δ is determined by a countable sequence of open sets there can be at most c G_δ 's, but for each $r \in \mathbf{R}$, $\mathbf{R} \setminus \{r\}$ is a dense G_δ , so there are exactly c dense G_δ 's.

Now let ω_c be the least ordinal with c predecessors and let $\{G_\alpha\}$ with $\alpha < \omega_c$ be the collection of all dense G_δ 's. Let g_1 be in G_1 , $g_1 \neq 0$, and suppose for $\alpha < \beta < \omega_c$ we have g_α in G_α . Think of \mathbf{R} as a vector space over the rationals and choose g_β in G_β with the condition that g_β is not in $\text{span}\{g_\alpha: \alpha < \beta\}$. This can be done since

$$\text{card}(\text{span}\{g_\alpha: \alpha < \beta\}) < c = \text{card}(G_\beta).$$

The set $A = \{g_\alpha: \alpha < \omega_c\}$ is clearly independent. To see that it is of 2nd category let's assume that it is of 1st category. A is therefore contained in a set F equal to the countable union of closed nowhere dense sets, hence the complement of F is a dense G_δ , say G_α . Thus g_α is in G_α and this contradiction concludes the argument.

Remarks:

1. Is there a set A such that $A + A$ contains an interval but $A \cdot A$ does not?
2. We get the same result for $\& = +$ or $-$ for a locally compact group $(G, +)$ since the category theorem is valid in locally compact Hausdorff spaces.
3. Does the image of a dense G_δ under a continuous function that is not constant on any interval contain a dense G_δ on some interval?
4. I would like to thank Joe Buhler for his valuable input.

REFERENCES

1. E. Hewitt and K. Ross, *Abstract harmonic analysis*, Springer Verlag 1963.
2. E. Hewitt and K. Stromberg, *Real and Abstract Analysis*, Springer Verlag 1965.
3. R. Mayer, *The Square of the Cantor set*, manuscript.
4. W. Rudin, *Fourier Analysis on Groups*, Wiley 1962.

3402 S. W. Corbett
Portland, OR 97201
steve@reed.edu

One can measure the importance of a scientific work by the number of earlier publications rendered superfluous by it.

—David Hilbert (1862–1943)

Mathematical Circles Revisited, Howard W. Eves,
Boston: Prindle, Weber and Schmidt, 1988.

UNSOLVED PROBLEMS

Edited by: Richard Guy and Richard Nowakowski

In this department the MONTHLY presents easily stated unsolved problems dealing with notions ordinarily encountered in undergraduate mathematics. Each problem should be accompanied by relevant references (if any are known to the author) and by a brief description of known partial or related results. Typescripts should be sent to Richard Guy, Department of Mathematics & Statistics, The University of Calgary, Alberta, Canada T2N 1N4.

Does the Möbius Function Determine Multiplicative Arithmetic?

D. Flath and A. Zulauf

Is the multiplication law on the positive integers uniquely determined by the values of the Möbius function and the property that multiplication respects order?

Let us be more precise. The Möbius function μ is defined on positive integers by the rule that $\mu(n)$ equals 0 if n is divisible by the square of a prime and $\mu(n)$ equals $(-1)^r$ if n is the product of r distinct primes. We define μ on free Abelian semigroups analogously, regarding the generators as primes.

MÖBIUS PROBLEM. Suppose that $A = \{a_1, a_2, a_3, \dots\}$ is a free Abelian semigroup, where a_1 is the unit element. Do the following two properties imply that $a_m a_n = a_{mn}$ for every m and n ?

1. $a < b$ implies $ac < bc$ for $a, b, c \in A$, where A has been given the linear order $a_1 < a_2 < a_3 < \dots$.
2. $\mu(a_n) = \mu(n)$ for every n .

The best way to understand the problem is to think of a_n as n and use the Möbius function and the order property to factor the first few positive integers, as follows.

To begin with, 1 is the unit.

Since $\mu(2) = -1$, 2 must be the product of an odd number of distinct primes. Since there are not three primes less than 2, 2 must be prime.

Similarly, $\mu(3) = -1$ implies that 3 is prime.

Next, $\mu(4) = 0$, so 4 is the smallest number divisible by the square of a prime; it must be the square of the smallest prime. Hence $4 = 2^2$.

Next, $\mu(5) = -1$ and since there are only two primes smaller than 5, 5 must be prime.

Since, $\mu(6) = 1$, 6 is the smallest number that is the product of distinct primes, so it must be the product of the smallest two. Thus $6 = 2 \cdot 3$.

Next, $\mu(7) = -1$. Can 7 be the product $2 \cdot 3 \cdot 5$? No, because $2 \cdot 3 \cdot 5 > 3 \cdot 5$ and no number less than 7 equals $3 \cdot 5$. Therefore 7 is prime.

Next, $\mu(8) = \mu(9) = 0$. A little logic shows that 8 and 9 must equal 2^3 and 3^2 , but there seems to be no way to determine which is which.

Go on to 10. Since $\mu(10) = 1$, it is easy to see that $10 = 2 \cdot 5$.

And so it goes. You can show that 11 is prime, that $12 = 2^2 \cdot 3$, and that 13 is prime quite easily. But 14 and 15 are $2 \cdot 7$ and $3 \cdot 5$ and there seems to be no way to tell which is which.

The further you go the more intricate the logic becomes. Some numbers factor easily, but more and more of them present problems. Let's skip ahead and see how the uncertainties about 8, 9, 14, and 15 can be resolved.

Let $Z(n) = \#\{m \leq n: \mu(m) = 0\}$. Then

$Z(2 \cdot 3 \cdot 5) = \#\{2^2 \cdot a, 3^2 \cdot b, 5^2: 1 \leq a \leq r, 1 \leq b \leq 3\} = r + 4 = 10$ or 11, where $r = 6$ or 7 according as $2 \cdot 7 >$ or $< 3 \cdot 5$. But $Z(n) = 10$ implies that $n = 27$, and $\mu(27) = 0 \neq \mu(2 \cdot 3 \cdot 5)$. Hence $Z(2 \cdot 3 \cdot 5) = 11$, $r = 7$, $2 \cdot 3 \cdot 5 \geq 29$, and $2 \cdot 7 < 3 \cdot 5$. Therefore $14 = 2 \cdot 7$ and $15 = 3 \cdot 5$.

Since $3 \cdot 11 > 2 \cdot 3 \cdot 5 \geq 29$ and $\mu(3 \cdot 11) = 1$, we have $3 \cdot 11 \geq 33$. Since $2^2 \cdot 3^2 > 3 \cdot 11 \geq 33$ and $\mu(2^2 \cdot 3^2) = 0$, we have $2^2 \cdot 3^2 \geq 36$, and hence that $Z(2^2 \cdot 3^2) \geq Z(36) = 13$. But

$$Z(2^2 \cdot 3^2) = \#\{2^2 \cdot a, 3^2 \cdot b, 5^2: 1 \leq a \leq s, 1 \leq b \leq 3\} = s + 4,$$

where $s = 8$ or 9 according as $2^3 > 3^2$ or $< 3^2$. Therefore $s = 9$, and so $2^3 = 8 < 3^2 = 9$.

One way to interpret what happens is that any confusion about two elements, say m and n , propagates and leads to confusion about $m \cdot k$ and $n \cdot k$ for $k = 2, 3, \dots$. Eventually $m \cdot k$ and $n \cdot k$ are sufficiently far apart for the confusion between them, and hence for that between m and n , to be resolvable. For instance, the confusion between 8 and 9 is resolved since $8 \cdot 4$ and $9 \cdot 4$ can be shown to be separated by $3 \cdot 11$. Going to higher numbers resolves uncertainties about factorizations again and again, but you seem to have to go farther and farther out to do it. Are all factorizations ultimately determined in this way? It is a question of just how much information is contained in the Möbius function, taken as an oracle.

The Möbius problem arose in 1979 from discussions between A. Zulauf and his doctoral student P. B. Braun. It was verified at the time that $a_{mn} = a_m a_n$ for all $mn \leq 74$ if $\mu(a_n) = \mu(n)$ for all $n \leq 240$, but the proof is lengthy and this comparatively insignificant result did not seem worth publishing. Meanwhile, P. B. Braun proposed and investigated a much stronger conjecture.

BRAUN'S CONJECTURE. Let $B = \{b_1, b_2, b_3, \dots\}$ be an infinite Abelian semi-group with the linear order $b_1 < b_2 < b_3 < \dots$ such that b_1 is the unit element and $a < b$ implies $ac < bc$ for $a, b, c \in B$. Define the Möbius function μ on B inductively by

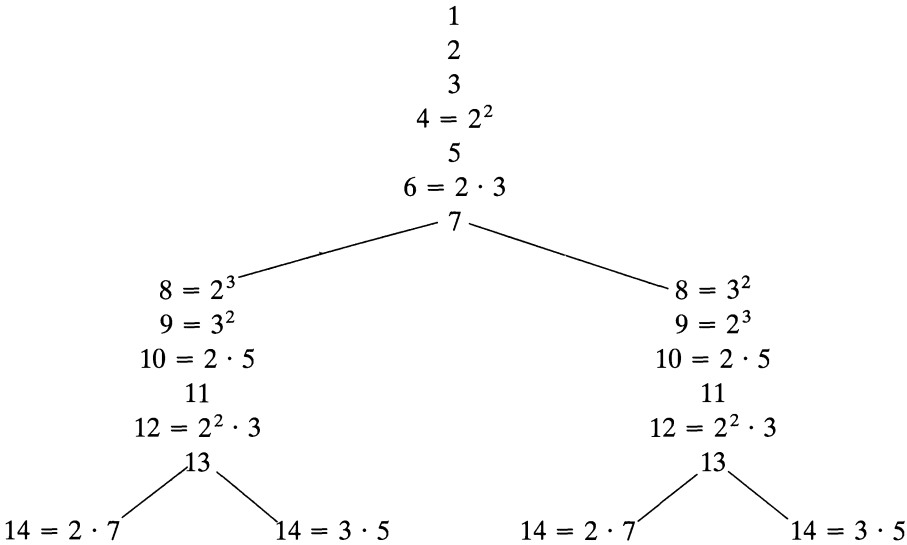
1. $\mu(b_1) = 1$
2. $\sum_{b_d | b_n} \mu(b_d) = 0$ for $n = 2, 3, 4, \dots$

Suppose that $\mu(b_n) = \mu(n)$ for all $n \geq 1$. Then $b_{mn} = b_m b_n$ for all $m, n \geq 1$.

The point here is that uniqueness of factorization is not assumed and, accordingly, the Möbius function is defined in a way that does not depend on each b_n being a unique product of powers of generators. Of course, if Braun's conjecture is

true then it follows that the fundamental theorem of arithmetic does in fact apply to B .

Here is a quick sketch of the way things work out for the first few positive integers. Considering the Möbius condition $\sum_{b_d|b_n} \mu(d) = 0$ for $n = 2, 3, 4, \dots$ in turn, and bearing in mind the ordering condition, one can build a tree of all possible factorizations on which each path from the root exhibits a string of factorizations of $b_1, b_2, b_3, \dots, b_n$ that meets all requirements. At $n = 14$ there will be four free branches, namely, writing n instead of b_n for brevity,



At $n = 21$ each free branch splits in two: $21 = 3 \cdot 7$ or $2 \cdot 11$. At $n = 22$ two of the eight free branches split in two, and it is at this stage that the possibility of non-unique factorization first occurs: there are two paths that have $22 = 3^3 = 5^2$ and, along the way, $8 = 3^2$, $9 = 2^3$, $20 = 2^2 \cdot 5$, $21 = 3 \cdot 7$. But in this case $7^2 > 5 \cdot 6 = 3 \cdot 10 > 3 \cdot 9 = 2 \cdot 12$, which would imply that $2 \cdot 11$ is less than all other composite elements exceeding 22, that $2 \cdot 11$ has unique factorization, and that $\mu(2 \cdot 11) = -\mu(1) - \mu(2) - \mu(11) = 1$ by the Möbius condition. But $n > 22$ and $\mu(n) = 1$ imply that $n \geq 26$. Hence 24 would be less than $2 \cdot 11$ and therefore irreducible, in contradiction of $\mu(24) = 0$. The possibility $22 = 3^3 = 5^2$ is thus eliminated, and the number of free branches at $n = 22$ reduces from ten to eight. More branches sprout at $n = 24$, 25 and 27, but at $n = 28$, the possibility $14 = 3 \cdot 5$ is eliminated, and at $n = 32$ the possibility $8 = 3^2$ is eliminated. This greatly reduces the number of free branches, and it has now been established that $b_{mn} = b_m b_n$ for all $mn \leq 20$ if $\mu(b_n) = \mu(n)$ for all $n \leq 32$. The uncertainty about 21 (and 22) is removed at $n = 46$, but by then many new uncertainties have been encountered, and one can but hope that all uncertainties are eventually resolved.

Department of Mathematics & Statistics
 University of South Alabama
 Mobile, AL 36688
 flath@mathstat.usouthal.edu

25, Hooker Avenue
 Hamilton, New Zealand

PROBLEMS AND SOLUTIONS

Edited by:
Richard T. Bumby, Fred Kochman and Douglas B. West

Proposed problems should be sent to the MONTHLY PROBLEMS address given on the inside front cover. Please include solutions, relevant references, etc. Three copies are requested.

Solutions of published problems should arrive before September 30, 1995 at the MONTHLY PROBLEMS address given on the inside front cover. Solutions should be typed with double spacing, including the problem number and the solver's name and mailing address. Two copies suffice. A self-addressed postcard or label should be included if an acknowledgement is desired.

An asterisk () after the number of a problem, or part of a problem, indicates that no solution is currently available. Partial solutions will be useful in such cases. Otherwise, the published solution is likely to be based on a solution which is complete and correct. Of course, an elegant partial solution or a method leading to a more general result is always useful and welcome. In addition, references to other appearances of MONTHLY problems or to solutions of these problems in the literature are also solicited.*

PROBLEMS

10445. *Proposed by Alan J. Gross, Medical University of South Carolina, Charleston, SC, and Hong Zhang, Indiana-Purdue University, Fort Wayne, IN.*

Note that $5^2 + 5 + 2 = 2^5$. Are there any other positive integers a and b with $a^b + a + b = b^a$?

10446. *Proposed by Hubert Kiechle, Technische Universität, Munich, Germany.*

Let $T = \{z : |z| = 1\}$ be the unit circle in the complex plane, and let w be a given nonzero complex number.

- (a) If $|w| \leq 2$, show that there are unique $z_1, z_2 \in T$ such that $w = z_1 + z_2$.
- (b) If $|w| > 2$, show that w can be written as a sum of $\lceil |w| \rceil$ elements of T .
- (c) Under what conditions will w be a unique sum of n elements of T .

10447. Proposed by Stephen C. Locke, Florida Atlantic University, Boca Raton, FL.

Consider a tournament in which every pair of teams play a match in which one of the two wins. Let L_0 be a listing of the teams in some order, and define successive $L_i, i = 1, 2, 3, \dots$ by repeated application of the following operation: if a team in the list L_i lost to the team immediately following it in the list, call that pair of teams a *switchable pair*; the order of one switchable pair is then reversed to give L_{i+1} . Note that this may increase the number of switchable pairs.

Prove that any such sequence of operations leads, in a finite number of steps, to a list in which every team defeated the team immediately following it in the list, so there are no switchable pairs.

10448. Proposed by Fu-Chuen Chang, National Sun Yat-sen University, Kaohsiung, Taiwan.

Fix a positive integer n . Let $x_i = \cos\left(\frac{(2i-1)\pi}{2n}\right)$ for $1 \leq i \leq n$, and $c_k = \frac{1}{n} \sum_{i=1}^n x_i^k$ for $k \in \mathbb{N}$. Show that

$$c_k = \begin{cases} 0 & k = 1, 3, \dots, 2n-1 \\ \binom{k}{k/2} 2^{-k} & k = 0, 2, \dots, 2n-2. \end{cases}$$

10449. Proposed by Frank Schmidt, Arlington, VA.

For which n can the symmetric group S_n be generated by two *conjugate* permutations?

10450. Proposed by Kenneth S. Williams, Carleton University, Ottawa, Ontario, Canada and Blair K. Spearman, Okanagan University College, Kelowna, B. C., Canada.

Let K be a quartic extension of the field \mathbb{Q} of rational numbers. K is called a *pure* extension of \mathbb{Q} if there is an integer l such that $K = \mathbb{Q}(\sqrt[4]{l})$, and K is called a *bicyclic* extension of \mathbb{Q} if there exist integers m and n such that $K = \mathbb{Q}(\sqrt{m}, \sqrt{n})$. Determine all quartic extensions that are both pure and bicyclic.

10451. Proposed by Joaquín Gómez Rey, I. B. “Luis Buñuel”, Alcorcón (Madrid), Spain.

In the story below, m, n and r are integers with $0 \leq r \leq m \leq n$.

“Once upon a time, there lived a miserly king who had m gold coins and n silver ones.

One day, he put $n + r$ coins in his right pocket, and the remaining $m - r$ coins in his left pocket. For the rest of his life, it was his pleasure, once each day, to take a coin at random from each pocket, privately admire the two coins, and then return each to the opposite pocket. In other words, he was a good king and lived happily in his castle for many years. In all those years, no one ever knew how many gold coins were in his left pocket on any particular day.”

Determine the most likely number of gold coins in his left pocket in the long run.

NOTES

(10447) This problem gives an alternative proof of Redei’s theorem that every tournament has a Hamiltonian path. (10448) The x_i are the roots of the Chebyshev polynomial $T_n(x)$. See Theodore J. Rivlin, *Chebyshev Polynomials*, Wiley, 1990, for more information about these polynomials. (10451) The process described here is an example of a “Pólya Urn Scheme”. It is related to the use of these processes to model diffusion. In W. Feller, *Introduction to Probability Theory*, vol. I, two such models are described. The process described here

is more similar to the Bernoulli-Laplace model, than to the Ehrenfest model, but Feller's exposition of both models may be helpful. This material is found in many places in the book, but may be easily located by using the index.

SOLUTIONS

Finitely Many Primes in Every Translate

10208 [1992, 266]. *Proposed by Solomon W. Golomb, University of Southern California, Los Angeles, CA.*

Let $1 \leq a_1 < a_2 < a_3 < \dots$ be an increasing sequence of positive integers.

- (a) Is there such a sequence $\{a_k\}$ having the property that, for all integers n (positive, negative, or zero), $\{a_k + n\}$ contains only finitely many primes?
- (b) Is there such a sequence $\{a_k\}$ and a constant $B > 0$ having the property that $\{a_k + n\}$ contains no more than B primes for every integer n ?

Solution (part a only) by Kevin Ford (student), University of Illinois, Urbana IL. The answer to part (a) is yes. Take $a_k = ((2k)!)! + k!$. If $|n| \geq 2$ and $k \geq |n|$, then $a_k > 2|n|$ and $|n|$ is a proper divisor of $a_k + n$. Hence, $\{a_k + n\}$ contains at most $|n| - 1$ primes when $|n| \geq 2$. Each of the sequences $\{a_k\}$ and $\{a_k + 1\}$ contain at most one prime since, for $k \geq 2$, k is a proper divisor of a_k and $k! + 1$ is a proper divisor of $a_k + 1$. If $k \geq 3$, then $k! - 1$ is a proper divisor of $a_k - 1$ and thus $\{a_k - 1\}$ contains at most two primes.

Editorial comment. Several readers observed that the fact that no answer was known to (b) demonstrated that the answer to part (a) was “yes”, but these solvers also went on to provide examples of sequences with this property. No solution to (b) was found. Kevin Ford, Robert High, Gerry Myerson, and the proposer all noted that a negative answer to (b) would result if the well-known “Prime k -tuples Conjecture” were true. This conjecture (given as **A9** in Richard K. Guy, *Unsolved Problems in Number Theory*, Springer-Verlag, 1981, p. 15) states that if $\{k_i\}$ is a finite set of integers and for every prime p the set of residues mod p do not form a complete residue system, then, for infinitely many values of n , the set $\{k_i + n\}$ consists entirely of primes. Following is Kevin Ford's proof that the Prime k -tuples Conjecture implies there does not exist a sequence $\{a_k\}$ and $B > 0$ with the desired properties.

First observe that the non-existence of such a sequence is equivalent to $f_{\{a_k\}}(L) \rightarrow \infty$ as $L \rightarrow \infty$, for every increasing sequence of positive integers, where

$$f_{\{a_k\}}(L) = \max_n (\text{number of primes in } \{a_1 + n, \dots, a_L + n\}).$$

For a sequence $\{a_k\}$ and integer $L \geq 1$, set $A_L = \{a_1, \dots, a_L\}$. Some residue class modulo 2 contains at most $L/2$ elements of A_L . Remove those from the set. Some residue class modulo 3 contains at most $1/3$ of the remaining elements. Remove these from the set. Continuing this process of removing one residue class modulo q for all primes $q \leq L$, we obtain a subset of A_L , say $\{z_1, \dots, z_K\}$, having the property that for every prime p , there is an integer m such that $z_i \not\equiv m \pmod{p}$, for every i . The Prime k -tuples Conjecture then implies that, for infinitely many integers n , the numbers $z_1 + n, \dots, z_K + n$ are all prime. It follows that

$$f_{\{a_k\}}(L) \geq K \geq L \prod \left(1 - \frac{1}{p}\right) > \frac{cL}{\log L},$$

where $c > 0$ is a constant independent of L and the product is over all primes less than or equal to L . Therefore, $\lim_{L \rightarrow \infty} f(L) = \infty$, as needed.

Solved also by T. Callahan, R. J. Chapman (U. K.), W. T. Gan (student, U. K.), J. W. Grossman, R. High, N. Komanda, J. H. Lindsey II, O. P. Lossers (The Netherlands), R. Martin (student), G. Myerson (Australia), A. Nijenhuis, A. Riese, R. M. Robinson, K. A. Ross, E. R. Scheinerman, K. Stoop (Switzerland), G. Thompson, Western Maryland College Problems group, University of Wyoming Problem Circle, and the proposer.

A Trace Inequality

10234 [1992, 571]. *Proposed by Götz Trenkler, University of Dortmund, Dortmund, Germany.*

Let A and B be nonnegative definite Hermitian matrices such that $A - B$ is also nonnegative definite. Show that $\operatorname{tr}(A^2) \geq \operatorname{tr}(B^2)$.

Solution I by Andreas Müller, Bures-sur-Yvette, France. The following proof shows that it suffices to assume only that $A + B$ and $A - B$ are nonnegative definite.

We regard A and B as selfadjoint operators on the Hilbert space $H = \mathbb{C}^n$. Since $A + B$ is hermitian and nonnegative definite, we can choose an orthogonal basis of eigenvectors of $A + B$, $e_i \in H$, $1 \leq i \leq n$. The eigenvalues λ_i (with $(A + B)e_i = \lambda_i e_i$) are all nonnegative. Then also

$$\langle e_i, (A - B)(A + B)e_i \rangle = \lambda_i \langle e_i, (A - B)e_i \rangle \geq 0$$

since $A - B$ is nonnegative definite. The trace of $(A - B)(A + B)$ is the sum of these terms, hence nonnegative. But,

$$\begin{aligned} (A - B)(A + B) &= A^2 - B^2 + AB - BA \\ 0 &\leq \operatorname{tr}((A - B)(A + B)) = \operatorname{tr}(A^2) - \operatorname{tr}(B^2) + \operatorname{tr}([A, B]) \\ &= \operatorname{tr}(A^2) - \operatorname{tr}(B^2) \end{aligned}$$

since $\operatorname{tr}(AB) = \operatorname{tr}(BA)$.

The claim and its proof remain valid for nonnegative operators of trace class in any separable Hilbert space. Trace class operators are compact. Hence the spectral theorem for hermitian compact operators guarantees the existence of a basis of eigenvectors for $A + B$. Also, the fact that the trace class operators form an ideal ensures that all the operators whose traces are needed in the above proof are of trace class.

Solution II by Thomas H. Foregger, AT & T Bell Laboratories, Warren, NJ. We have $\operatorname{tr}(A^2) - \operatorname{tr}(B^2) = \operatorname{tr}((A - B)A) + \operatorname{tr}(B(A - B))$, so the result follows from

Lemma. *If X and Y are nonnegative definite hermitian matrices, then $\operatorname{tr}(XY) \geq 0$.*

Proof. For such X and Y , there exist U and V such that $X = UU^*$ and $Y = VV^*$. Hence,

$$\operatorname{tr}(XY) = \operatorname{tr}(UU^*VV^*) = \operatorname{tr}(U^*VV^*U) = \operatorname{tr}((U^*V)(U^*V)^*) \geq 0.$$

Solution III by Duane W. Bailey, Amherst College, Amherst, MA. If X is an n by n hermitian matrix, let $\lambda_1(X) \leq \lambda_2(X) \leq \dots \leq \lambda_n(X)$ denote its eigenvalues arranged in order. Then, since $A - B$ is nonnegative definite, Corollary 4.3.3 on p. 182 of Roger A. Horn & Charles R. Johnson, *Matrix Analysis*, Cambridge, 1985 gives

$$\lambda_k(B) \leq \lambda_k(B + (A - B)) = \lambda_k(A) \quad \text{for } k = 1, 2, \dots, n.$$

If we further assume that B is nonnegative definite, it follows that A is also nonnegative definite and

$$\operatorname{tr}(B^2) = \sum_{k=1}^n \lambda_k(B)^2 \leq \sum_{k=1}^n \lambda_k(A)^2 = \operatorname{tr}(A^2).$$

Editorial comment. Many solvers recognized that the result could be obtained in different ways and gave two proofs. Pei Yuan Wu noted that Solution III easily gives that $\text{tr}(A^p) \geq \text{tr}(B^p)$ for any nonnegative number p , and that the same conclusion for $0 \leq p \leq 2$ can be obtained from Solution II since $A^{p/2} - B^{p/2}$ will be nonnegative definite for these p (see Theorem 3 of N. N. Chan & Man Kam Kwong, “Hermitian matrix inequalities and a conjecture”, this MONTHLY, 92 (1985), 533–541). Dennis I. Merino gave a converse to the lemma of Solution II: A given matrix A is nonnegative definite if and only if $\text{tr}(XA) \geq 0$ for every nonnegative definite matrix X .

Solved also by K. V. Bhagwat (India), D. Callan, R. J. Chapman (U. K.), J. Dai & Q. Luo (China), E. A. Herman, R. H. Jeurissen (The Netherlands), N. Kang (student, Korea), M. K. Kinyon, O. Krafft (Germany), G. Letac (France), D. I. Merino, J. M. Monier (France), A. Nijenhuis, I. Olkin, H. Özdemir (Turkey), M. Qian, E. T. Wong, P. Y. Wu (China), University of Wyoming Problem Circle, and the proposer.

Area of a Roulette

10254 [1992, 782]. *Proposed by E. Ehrhart, Université de Strasbourg, Strasbourg, France.*

The curve traced out by a fixed point of a closed convex curve as that curve rolls without slipping along a second curve will be called a “roulette”. Let S be the area of one arch of a roulette traced out by an ellipse of area s rolling on a straight line. Prove or disprove that $S \geq 3s$, with equality only if the ellipse is a circle.

Solution by Murray S. Klamkin, University of Alberta, Edmonton, Alberta, Canada. We will show that the inequality is equivalent to $(a - b)(a - 2b) \geq 0$ where a and b are the major and minor semi-axes of the ellipse, respectively. Consequently, there will be equality if $a = b$ or $a = 2b$. There will be strict inequality only if $a > 2b$.

Recall that the *pedal* of a given curve with respect to a point P is the locus of the foot of the perpendicular from P to a variable tangent line to the curve. The desired result follows from the following results of Steiner that can be found in B. Williamson, *The Integral Calculus*, Longmans, Green and Co., London, 1941, 201–203.

(A) When a closed curve rolls on a straight line, the area between the line and the roulette generated in a complete revolution by any point on the rolling curve is double the area of the pedal of the rolling curve, this pedal being taken with respect to the generating point.

(B) The area of the pedal of an ellipse of semiaxes a and b with respect to any point P is given by $\pi(a^2 + b^2 + |OP|^2)/2$, where O is the center of the ellipse.

In the interest of simplicity, these theorems have been stated only when P lies on the curve. This is not an essential restriction.

Clearly, the minimum of S for P on the ellipse occurs for $|OP| = b$. Hence, $S \geq 3s$ is equivalent to

$$\pi(a^2 + 2b^2) \geq 3\pi ab \quad \text{or} \quad (a - b)(a - 2b) \geq 0.$$

Editorial comment. The other solvers were able to work through the Calculus without references, but as one of them said: “. . . I hope some are more elegant in the way they prove the result; I just ground out the integral. . .”.

The work leading to the formulation of this problem can be found in E. Ehrhart, “Les roulettes d’ellipses, *L’Ouvert* 62 (1991), 43–45.

Other references to Steiner’s theorem found by the editors are E. Goursat, *A Course in Mathematical Analysis*, Vol. I, Dover, 1959, where it is problem 23 (with hints) on p. 207, and J. Edwards, *A Treatise on the Integral Calculus*, Chelsea, 1955, article 673, pp. 696–697, which refers back to W. H. Besant, *Tract on Roulettes and Glisettes*, 1870 (and not to Steiner).

Richard Holzsager suggested that it would be interesting to find the convex curve C and point P on C which gives the minimum ratio of the area of the roulette to the area of the

curve. He conjectured that C is given by the arc of the epicycloid $x = 3 \cos \theta - \cos 3\theta$, $y = 3 \sin \theta - \sin 3\theta$ with its endpoints $(\pm 2, 0)$ connected by a line segment. The point P is $(0, 0)$. For this curve, the ratio can be calculated to be $8/3$ while the above results show that when C is an ellipse, the smallest value is $2\sqrt{2}$.

Solved also by J. Anglesio (France), M. V. Bjelica (Yugoslavia), and R. Holzsager. One incorrect solution was received.

The Case of Horological Interchangeability

10260 [1992, 873]. *Proposed by Gerald Weinstein, The City College, CUNY, New York, NY.*

A man has a bizarre watch with indistinguishable hands. An act of violence, taking place sometime between midnight and the following noon, simultaneously kills him and stops his watch. Is it always possible to determine the time of death uniquely from this information if:

- (a) the watch has hour, minute and second hands?
- (b) the watch has only hour and minute hands?

Solution by John H. Lindsey II, Ft. Myers, FL. Let the positions of the hour, minute, and second hands as a fraction of the way around the dial from 12 be given by x, y, z , respectively. Then $0 \leq x, y, z < 1$, and there are integers $0 \leq m < 12$ and $0 \leq n < 60$ such that $12x = m + y$ and $60y = n + z$. An ambiguity means that analogous equations with integers i, j in place of m, n hold after some nonidentity permutation of x, y, z .

First consider the transposition (x, y) . Using the additional equation $12y = i + x$, we have $144y = 12(i + x) = 12i + m + y$. Thus $y = (12i + m)/143$, and hence $x = (m + y)/12 = (12m + i)/143$. If there is no second hand, any choice of $i \neq m$ will yield an ambiguous time. Thus the answer to (b) is no. If there is a second hand, however, we have $y - x = \frac{n+z}{60} - \frac{j+z}{60} = \frac{n-j}{60}$, and also $y - x = \frac{12i+m}{143} - \frac{12m+i}{143} = \frac{11(i-m)}{143} = \frac{i-m}{13}$. Since 13 and 60 are relatively prime, this forces $y - x$ to be an integer. Hence $x = y$ and there is no ambiguity.

Since the hour hand determines the time by itself, the transposition (y, z) cannot yield an ambiguity. The transposition (x, z) leaves the minute hand unchanged. Since the minute hand determines the second hand, this forces $x = z$ and there is no ambiguity.

Finally, we consider 3-cycles. As (x, y, z) and (x, z, y) are inverses, we need only consider one of them. Suppose $12z = i + x$ and $60x = j + y$. Then $j + y = 5(12x) = 5(m + y)$, and hence $4y = j - 5m$. But now $n + z = 60y = 15(j - 5m) \in \mathbb{Z}$. This requires $z = 0$. Since z is the position of the hour hand in the second reading, this requires $x = y = z = 0$. If the event was *strictly* between midnight and noon, then the answer to (a) is yes.

Editorial comment. Stan Wagon pointed out that part (b) previously appeared as problem E1571 [1963, 330; 1964, 91]. This two-handed case also appears in Thomas Szirtes, "On the problem of the interchangeable clock hands," *Journal of Recreational Mathematics* 8 (1975–76), 159–168 and an approach to a generalization is sketched in Karel A. Post, "Letter to the editor", *Journal of Recreational Mathematics* 11 (1978–79), 41. These articles contain suggestions that the problem is quite old. This is confirmed by its appearance as problem 48 in H. E. Dudeney, *536 Puzzles and Curious Problems* (Martin Gardner, ed.), Scribners, 1967, which is a reprint of earlier collections of problems. It also appears as chapter 143 of Joe Roberts, *Lure of the Integers*, MAA, 1992.

Solved also by J. Andraos (Canada), M. Bowron, R. J. Chapman (U. K.), M. P. Eisner, D. L. Grant & M. Heggie (Canada), I. Kastanas, O. P. Lossers (The Netherlands), W. D. McIntosh, N. Passell, S. Paul & J. Hess, R. E. Prather, K. Y. Tsang, E. A. Weinstein, Anchorage Math Solutions Group, GCHQ Problem Solving Group (U. K.), Trinity University Problem Group, and the proposer. Three other solvers correctly solved part (b) only, and one incorrect solution was received.

Identities for the Catalan Generating Function

10264 [1992, 874]. *Proposed by L. W. Shapiro, Howard University, Washington DC, and D. G. Rogers, Australian National University, Canberra, Australia.*

Let $C_n = 1/(n+1)\binom{2n}{n}$ for $n \in \mathbb{N}$ and form the generating function

$$C(x) = \sum_{n \geq 0} C_n x^n.$$

Establish the identities

$$(a) \sum_{n \geq 0} (n+1)x^n C(x)^{2n+2} = \sum_{m \geq 0} (4x)^m.$$

$$(b) \sum_{n \geq 0} (2n+1)x^n C(x)^{2n+1} = \sum_{m \geq 0} (4x)^m.$$

Solution by J. C. Binz, University of Bern, Bern, Switzerland. The generating function for the Catalan numbers C_n is well known to be $C(x) = \frac{1-\sqrt{1-4x}}{2x}$. We use the easily verified relations

$$xC(x)^2 = C(x) - 1 \quad \text{and} \quad \frac{C(x)}{2 - C(x)} = \frac{1}{\sqrt{1-4x}}.$$

With $y = xC(x)^2$, we obtain

$$\begin{aligned} \sum_{n \geq 0} (n+1)x^n C(x)^{2n+2} &= C(x)^2 \sum_{n \geq 0} (n+1)y^n = \frac{C(x)^2}{(1-y)^2} \\ &= \left(\frac{C(x)}{2 - C(x)} \right)^2 = \frac{1}{1-4x} = \sum_{m \geq 0} (4x)^m, \end{aligned}$$

and similarly,

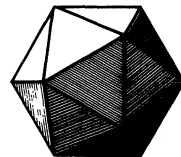
$$\begin{aligned} \sum_{n \geq 0} (2n+1)x^n C(x)^{2n+1} &= C(x) \sum_{n \geq 0} (2n+1)y^n = \frac{C(x)(1+y)}{(1-y)^2} \\ &= \left(\frac{C(x)}{2 - C(x)} \right)^2 = \frac{1}{1-4x} = \sum_{m \geq 0} (4x)^m. \end{aligned}$$

Editorial comment. Combinatorial interpretations of the formulas were submitted by David Callan and Renzo Sprugnoli.

Solved also by S.-J. Bang (Korea), J. L. Bryant, D. Callan, R. J. Chapman (U. K.), J. L. Drost, S. Getu, I. Kasantas, P. Kirschenhofer (Austria), M. S. Klamkin (Canada), I. I. Kotlarski, Y. H. Kwong, O. P. Lossers (The Netherlands), R. Martin (student), C. R. Pranesachar (India), H. Prodinger (Austria), R. Richberg (Germany), E. Schmeichel, R. Sprugnoli (Italy), D. B. Tyler, F.-Z. Zhao, GCHQ Problem Solving Group (U. K.), Western Maryland College Problems group, University of Wyoming Problem Circle, and the proposers. One incomplete solution was received.

Collaborating editors: David F. Appleyard, Paul T. Bateman, Bruce C. Berndt, Duane M. Broline, Barry W. Brunson, Frank S. Cater, Gulbank D. Chakerian, Underwood Dudley, Gerald A. Edgar, Michael A. Filaseta, Ira M. Gessel, Richard A. Gibbs, Jerrold R. Griggs, Douglas A. Hensley, John R. Isbell, Mourad E. H. Ismail, Murray Klamkin, Daniel J. Kleitman, Frederick W. Luttman, Frank B. Miles, Richard Pfeifer, Stephen L. Portnoy, J. O. Shallit, John Henry Steelman, Kenneth B. Stolarsky, David E. Tepper, Douglas B. Tyler, Daniel Ullman, and William E. Watkins.

The American Mathematical Monthly



Volume 102, Number 5 / MAY 1995



Hermann Weyl
(See p. 453.)

NOTICE TO AUTHORS

The *Monthly* publishes articles, notes, and other features about mathematics and the profession. The readership of the *Monthly* is intended to include everybody who is mathematically inclined, including of course professional mathematicians and students of mathematics at all collegiate levels. While no single article or feature is likely to appeal to everyone, material should interest and be accessible to a large number of readers. This is the most important criterion for acceptance.

Articles may be expositions of old results or presentations of new ones. They may concern all of mathematics or one small area, a broad development or a single application, historical reminiscences or one important event. While some articles may contain the author's new research, the novelty of material and generality of the results is far less important than the clarity of exposition and general interest. Discussing one illuminating case of a well known result is far better than providing all the details of an obscure but new proposition. Articles in the *Monthly* are supposed to inform and to entertain; they are meant to be read rather than archived.

Notes are short and possibly informal articles. A note may concern a clever new proof of an old theorem, a novel way to present tired material, or a lively discussion of a philosophical (but still mathematical) issue. Also, any topic is suitable, so long as it is related to mathematics. Because a note is short, the first few sentences are the most important part: They should explain the purpose and invite the reader in. Photographs or diagrams often will attract the reader's attention.

All articles and notes should be sent to the editor:

JOHN EWING
Department of Mathematics
Indiana University
Bloomington, IN 47405

Please send 3 copies, typewritten on only one side of the paper. Illustrations should be carefully drawn on separate sheets of paper in black ink; the original should be without lettering and two copies should have appropriate captions and lettering indicated.

Proposed problems or solutions should be sent to:

RICHARD BUMBY,
P.O. Box 10971
New Brunswick, NJ 08906-0971.

Please send 2 copies of all material, typewritten if possible.

Letters to the Editor, both for publication and for private reading, should be sent to the Editor at the address given above. Comments, including criticisms, are welcome, as are all suggestions for making the *Monthly* a lively, entertaining, and informative journal.

EDITOR:

JOHN H. EWING

ASSOCIATE EDITORS:

PETER BORWEIN	FRED KOCHMAN
RICHARD BUMBY	CATHERINE MCGEOCH
DENNIS DETURCK	RICHARD NOWAKOWSKI
UNDERWOOD DUDLEY	ARNOLD OSTELEE
JOHN DUNCAN	LEE RUBEL
JOAN FERRINI-MUNDY	ABE SHENITZER
JOSEPH GALLIAN	LYNN STEEN
STEVEN GALOVICH	STAN WAGON
RICHARD GUY	DOUGLAS WEST
DARRELL HAILE	HERBERT WILF
PAUL HALMOS	SANDY ZABELL
JOAN HUTCHINSON	PAUL ZORN

EDITORIAL ASSISTANT:

MISTY CUMMINGS

STAFF ARTIST:

MIKE CAGLE

QUOTE MASTER:

MARK WOODARD

Reprint permission:

MARCIA P. SWARD, Executive Director

Advertising Correspondence:

Ms. ELAINE PEDREIRA, Advertising Manager

Subscription correspondence, change of address, and other inquiries:

Membership / Subscriptions Department

All at the address:

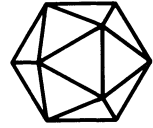
The Mathematical Association of America
1529 Eighteenth Street, N.W.
Washington, DC 20036.

Microfilm Editions: University Microfilms International, Serial Bid coordinator, 300 North Zeeb Road, Ann Arbor, MI 48106.

The AMERICAN MATHEMATICAL MONTHLY (ISSN 0002-9890) is published monthly except bimonthly June-July and August-September by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, DC 20036 and Montpelier, VT. Copyrighted by the Mathematical Association of America (Incorporated), 1995, including rights to this journal issue as a whole and, except where otherwise noted, rights to each individual contribution. General permission is granted to Institutional Members of the MAA for noncommercial reproduction in limited quantities of individual articles (in whole or in part) provided a complete reference is made to the source. Second class postage paid at Washington, DC, and additional mailing offices. **Postmaster:** Send address changes to the American Mathematical Monthly, Membership / Subscription Department, MAA, 1529 Eighteenth Street, N.W., Washington, DC, 20036-1385.

The American Mathematical Monthly

Volume 102 Number 5 / MAY 1995
(ISSN 0002-9890)



Contents

ARTICLES

Why Did George Green Write His Essay of 1828 on Electricity
and Magnetism?/I. GRATTAN-GUINNESS 387

Communicating Mathematics: Useful Ideas from Computer
Science/CHARLES WELLS 397

Order and Chaos on Your Desk/SUSAN BASSEIN 409

On the Geometry of Halley's Method/T. R. SCAVO and J. B. THOO 417

The Binary Expansion of $\frac{1}{p}$ /A. R. MEIJER 427

Pick's Formula via the Weierstrass \wp -Function/RICARDO DIAZ and SINAI
ROBINS 431

FEATURES

COMMENTS 386

NOTES

Permutations as Products of Transpositions/
GEORGE MACKIW 438

Congruences Relating the Order of a Group to the Number
of Conjugacy Classes/BJORN POONEN 440

The Color Invariant for Knots and Links/PETER ANDERSSON 442

THE COMPUTER SCIENCE SAMPLER

Veni, Divisi, Vici/CATHERINE C. MCGEOCH 449

THE EVOLUTION OF...

Part I. Topology and Abstract Algebra as Two Roads of Mathematical
Comprehension/HERMANN WEYL 453

THE AUTHORS 461

PROBLEMS AND SOLUTIONS 463

REVIEWS

*The World's Most Famous Math Problem (The Proof of Fermat's Last
Theorem and Other Mathematical Mysteries).* By Marilyn vos Savant/
NIGEL BOSTON and ANDREW GRANVILLE 470

TELEGRAPHIC REVIEWS 474

Why Did George Green Write His Essay of 1828 on Electricity and Magnetism?

I. Grattan-Guinness

1. HONOUR TO GREEN. Among the centenaries of mathematicians and scientists celebrated in 1993, perhaps the most remarkable was the bicentenary of the birth of a professional miller and part-time mathematician, one George Green (1793–1841) of Sneinton, then near Nottingham. Among other achievements, he was the creator of theorems and functions now named after him which make him a principal contributor to potential theory and to its applications in mechanics and mathematical physics.

During the week corresponding to that of his birth (which occurred on 14 July) various events took place. A three-day conference was held at the University of Nottingham, mainly on the use of his work in modern mathematics and physics. It included a visit to the mill at Sneinton, which had been restored and opened as a science centre in 1985. The next day a stained glass window was dedicated at the Gonville and Caius College, Cambridge, where he was resident from 1833 to 1837 as an extremely mature student, and (only) for the winter of 1839–1840 as a Fellow. Finally, on Friday 16 July a meeting was held at the Royal Society of London on his life and work and the modern importance of the latter. It was followed by a quite exceptional event: the unveiling of a plaque in his memory in the floor of the nave of Westminster Abbey, close to the tomb of Isaac Newton and to the plaques for his first publicist Lord Kelvin, Michael Faraday and Clerk Maxwell.

These events had been preceded by the publication in May of an excellent biography of Green (Cannell 1993)—a daunting task to write, as his life is so obscure (for example, no surviving likeness or portrait has ever been found, and his manuscripts seem to have been destroyed). It is clear, though, that in virtual isolation at Sneinton he taught himself Continental mathematics, and produced first-class research work. It was published in 1828 (his 38th year) as a 72-page *Essay on the Mathematical Analysis of Electricity and Magnetism* (Green 1828), put out at his own expense with the help of a subscription list. Largely ignored during the author's lifetime, it has since been reprinted no less than seven times and translated into German. How and why was it created?

2. THREE STRANDS IN 18TH-CENTURY MECHANICS. One of the most profound influences exercised by this fugitive work is that it raised both the status of potential theory in mathematics and the quality of the theorems that could be stated in it. Prior to this time three strands of thought in potential theory (as we now understand it) were active, though not necessarily with close links between them (Todhunter 1873, vol. 1).

The most significant strand was the attraction of spheroids to an external point. Isaac Newton had found various special properties in the *Principia*, in his synthetic

style: they were extended from the 1770s onwards by P. S. Laplace and A. M. Legendre using analytical methods, especially the Legendre functions and surface and zonal harmonics.

Another line came from Alexis Clairaut on the Continent from the 1730s (with some contributions from Colin MacLaurin in Britain soon afterwards), where properties of equipotential surfaces were studied; this work laid stress on the exact differential of a function of several variables, and assisted in the birth soon afterwards of the full partial differential calculus. With d'Alembert, some aspects of Euler's work, and especially J. L. Lagrange, variational mechanics was developed, in which force and velocity potentials were often used in the formation of differential equations.

A third strand grew out of Daniel Bernoulli's *Hydrodynamica* (1738), where considerations of 'ascensus actualis et potentialis' led to conservation of energy as a basis for (much) mechanics; his notions were to end up in the next century as kinetic and potential energy respectively, although with substantial changes in conception in which potential theory was to play a role.

In addition, an isolated contribution came from Lagrange in 1762. While pondering ways of solving the equations for the propagation of sound in three dimensions, he formed volume integrals of the solution in each co-ordinate direction, integrated them by parts to create surface integrals, and then added up the resulting equations to obtain a simpler differential equation to integrate. A clever but ad hoc manoeuvre, it had little influence even upon its distinguished innovator; but it was closer to the way ahead pursued in the early 19th century than the strands just mentioned.

3. POISSON AND THE APPEARANCE OF DIVERGENCE THEOREMS. Enter Siméon-Denis Poisson (1781–1840), the leading supporting actor in this drama, student and then professor and graduation examiner at the *Ecole Polytechnique*, devout follower of Laplace and Lagrange in mathematical methods and physical modelling. Poisson inaugurated mathematical electrostatics (I shall use the word 'electricity' of the time) in two papers (Poisson 1812, 1814) published by the Paris Academy of Sciences in which he analysed arrangements studied experimentally by C. A. Coulomb 30 years previously; equilibrium on a charged spheroid, and between two spheres. The principal mathematical exercise was to modify Legendre functions and related potential theory to fit the assumptions made about the phenomena (Grattan-Guinness 1990, 496–513).

In a short paper written soon after these two, (Poisson 1813) rectified an important oversight of his masters when he pointed out that the differential equation governing the potential V to a body A relative to an interior point M was not Laplace's equation

$$\Delta V = 0 \text{ (with '}\Delta\text{' as the Laplacian operator) but } \Delta V = -4\pi\rho, \quad (1)$$

where ρ was the density of material at M . He might have got this insight from his recent work on electricity; another strong candidate is a paper of that year on the attraction of spheroids by Carl Friedrich Gauss, which contained a result which in vectors reads

$$\int_S d\mathbf{s} \cdot \mathbf{r} = 0 \quad \text{or} \quad = -4\pi, \quad \text{where } \mathbf{r} = B\mathbf{M} \quad (2)$$

and B is an arbitrary point in A , according as M is outside *or* inside the surface S of A . Neither man dealt with the case where M is *on* S , when -2π obtains in (1)₂ (Grattan-Guinness 1990, 418–424).

Twelve years later Poisson came to the Academy of Sciences with another pair of papers, this time analysing magnetism (Poisson 1826a–b). Taking a magnetic body A to be composed of discrete ‘magnetic elements’ D , he set to zero certain surface integrals over D expressing internal equilibrium, and wrote down volume integrals to state the components of attraction of A to an external point M relative to his imposed rectangular co-ordinate system (x, y, z) . The second part of the first paper dealt with a ‘simplification of the preceding formulae’; integrating these integrals by parts with respect to (say) z led him to convert the volume integral to an integral over the surface S of A . I write his finding in the form

$$\iiint_A H_z(x, y, z) \, dx \, dy \, dz = \iint_S H(x_S, y_S, z_S) \cos n \, dS, \quad (3)$$

where H was the function expressing the components of magnetic attraction, and n was the angle between the z -axis and the normal at the point (x_S, y_S, z_S) of S . Adding this formula to its brothers for the x - and y -directions gave him the first general divergence theorem in mathematics; imitating the notation of (3), it can be written

$$\begin{aligned} & \iiint_A [F_x + G_y + H_z](x, y, z) \, dx \, dy \, dz \\ &= \iint_S [F \cos l + G \cos m + H \cos n](x_S, y_S, z_S) \, dS. \end{aligned} \quad (4)$$

He modified it for the case when M was inside A by the manner of his proof of (1)₂, and found a new term involving a factor $-4\pi/3$.

Poisson knew that his result was not restricted to convex bodies (a sum of integrals of the form (3)₂ is required as the z -axis goes in and out of A), nor to magnetism. But he saw it simply as a convenience; triple integrals are replaced by double integrals (Grattan-Guinness 1990, 948–953). This point will be crucial for Green, as we shall see the next section.

In a third paper, published by the Academy in the *Mémoires* (Poisson 1827), he analysed the process of magnetisation in moving bodies. A most complicated analysis used Legendre functions once again; but an important detail was his recollection of his equation (1)₂ for interior points, and first presentation of the version with $-2\pi\rho$ for surface points.

Surface integrals were enjoying a springtime in French mathematics at this time. For example, Adrien Marie Ampère had been studying electromagnetism and electrodynamics (his word) since 1820; his analysis made adroit use of both surface and line integrals, the latter arising naturally in connection with the attraction caused by current-bearing wires (Grattan-Guinness 1990, 941–961). One of his most remarkable results, published in 1826, was to show that Poisson’s basic formulae for magnetism could be restated in his own preferred conception, which saw magnetism as a special case of electricity and so replaced Poisson’s ‘magnetic elements’ with a tiny electrical solenoid (his word again).

Another source was Joseph Fourier’s pioneering work on heat diffusion, created in the mid 1800s, fully published only in the early 1820s, especially in his book *Théorie analytique de la chaleur* (1822), and then receiving much attention from the new generation of French mathematicians. In particular, around 1826 Jean Duhamel and Russian visitor Mikhail Ostrogradsky independently sought to justify mathematically Fourier’s use of trigonometric series solutions (Grattan-Guinness 1990, 1168–1176). Let f and g be two different special solutions for diffusion in a

body A , and consider $I := \iiint_A fg \, dV$. Integrating by parts through A led to a divergence theorem like Poisson's (4); and applying Fourier's external surface condition showed that in fact $I = 0$. Hence f and g were orthogonal over A , like the sine and cosine functions. We can see that this does not provide the justification sought; more to the point is the use again of surface integrals and a divergence theorem.

Although these integrals were making appearance, their presence in mathematics was still slight. Good evidence is provided by Augustin Louis Cauchy (1789–1857), former pupil of the *Ecole Polytechnique* (when Poisson was professor) and now professor there himself, inaugurating his revision of the calculus and mathematical analysis by his famous new approach with the theory of limits at the centre, emphasis laid upon continuity of functions. (Poisson and others there protested vigorously.) Above all, the derivative and integral were defined *separately*, so that the fundamental theorem of calculus became a proper theorem for the first time (Grattan-Guinness 1990, 707–804, including Cauchy's concurrent inauguration of complex-variable analysis). However, he never furnished a definition of either the line or the surface integral, although the required forms of definition would not have been hard to devise; they were too marginal to be worth such attention.

Then Green started thinking.

4. GREEN AND THE PLACE OF SURFACE INTEGRALS. Possible sources for Green's essay will be appraised in the next section; here its main contents are described. Pages are cited from the printing in the edition of his works (Green 1871).

After various preliminaries, the essay contains two roughly equal parts on electricity and on magnetism, in that order. These latter analyses draw heavily on various largely known integral expressions to state external and internal potentials (the latter maybe learnt from Poisson's (1)₂), and Legendre functions to express the potentials in analytical form. He extended various results due to Poisson, and considered some variant situations, such as when the spheres are connected by a wire (Whitrow 1984).

The chief novelties were presented in the 'general preliminary results' stated in the opening. First was the explicit specification of 'the potential function,' as he called it (1828, 9), and now named after him:

It only remains therefore to find a function V' which satisfies the partial differential equation, becomes equal to [a given function] \bar{V}' when [the point] p is upon the surface A , vanishes when p is at an infinite distance from A , and is besides such that none of its differential co-efficients shall be infinite when the point p is exterior to A .

(1828, 12: note the inadequate specification of $V'(\infty)$, and that the prime does not denote differentiation). This formulation anticipated in certain ways the 'Dirichlet principle', which was to assume such status in potential theory when its author began lecturing on the subject from 1839 in Berlin: a decade earlier he also was in Paris, but working on Fourier's heat theory, Cauchy's analysis, and number theory.

Secondly was Green's type of divergence theorem, expressed entirely within the rectangular co-ordinate system (x, y, z) rather than with surface differentials of Poisson's (4): for two 'continuous functions' $U(x, y, z)$ and $V(x, y, z)$

$$' \int dx dy dz U \delta V + \int d\sigma U (dV/dw) = \int dx dy dz V \delta U + \int d\sigma V (du/dw) ' \quad (5)$$

(1828, 23). I follow his use of ' δ ' for the Laplacian operator (an unusual symbol, perhaps required by the limitations of his printer's font box), ' $d\sigma$ ' for the element of the surface, all integrals stated with only one sign ' \int ' (unlike Poisson's use of multiple integral signs), and round brackets to indicate partial 'differential co-efficients' (Euler's practice, and name also, both of which Green followed). He modified his result for 'singularities' in U (or V) at points G by adding in terms of the form $-4\pi U(x_G, y_G, z_G)$ to the appropriate side of the equation (1828, 27), like Poisson's own modification; he may also have known of Poisson's equation (1)₂ from its reappearance in (Poisson 1827). I wonder at the import of the continuity imposed upon U and V , and the reference to 'singularities'; had he also been reading Cauchy on reforming the calculus?

Green had taken up a current research interest in mathematical physics in using volume and surface integrals to analyse electricity and magnetism; and with his insights he surpassed all contemporaries. This theorem (5), while similar in mathematical form to Poisson's (4), was understood at a far deeper level as physics (and also surpassed Gauss's (2) in generality). Whereas Poisson saw only simplification in his integral, Green recognised that *the importance of his own theorem lay in relating properties inside bodies to properties on their surfaces and vice versa*. He must have realised that theorems of this kind served for multiple integral calculus like the fundamental theorem of the calculus itself; hence the importance of integration by parts.

These insights doubtless led Green further to the novelty of his function V' in which conditions in a body and on its surface were imposed. Such functions were found for various cases with the help of his theorem; one of them followed it in its symmetrical form (Green 1828, 37–39), and launched what have become known as 'reciprocity relations'.

One may guess therefore, with some confidence, *that Poisson's first two papers on magnetism were the source of inspiration for Green's research, especially the divergence theorem* (4). Up to then Green had doubtless been learning mathematical skills and theories, but he had not found a *deep problem*: Poisson (unintentionally) provided this, in the form of an unexceptionable but somewhat limited use of Legendre functions to analyse the distribution of magnetism, and especially in a 'simplification' which held much deeper consequences than its author had realised.

5. SOURCES AND INFLUENCES. While it is possible to guess at Green's original motivation, his training in mathematics remains unknown. Among local figures, headmaster John Toplis (1774/1775–1857) would have been a crucial figure in forming the interests of his former pupil at Nottingham Grammar School: a deplorer of the state of British mathematics in the *Philosophical Magazine* in 1804, a translator of Lacroix there a year later, and of Book One of Laplace's *Mécanique céleste* in a book published in Nottingham in 1814. However, in 1819 he returned to his college (Queen's, Cambridge), and was *not* to be one of the subscribers. The only other likely supporter is Sir Edward Bromhead (1789–1855: so Green's senior by a mere four years), member with Charles Babbage and John Herschel of the Analytical Society at Cambridge in the mid 1810s, and a subscriber to the essay; but his letter of April 1828 acknowledging receipt of his presentation copy (Cannell 1993, 67) shows that he had *not* been aware of its contents before it arrived.

Green's access to literature is also little understood. In his essay he cited as mathematical sources Laplace's *Mécanique céleste*, Book 3 (1799) for Legendre functions, Fourier's *Théorie analytique de la chaleur*, and of course, Poisson's three

papers on magnetism and the two on electricity; Boit's *Traité de physique* (1816) was used for information on Coulomb's experiments. A passing reference (1828, 103) to Lagrange's follower L. F. A. Arbogast shows his familiarity with some of the current French operator techniques. A sentence in his introduction comparing Fourier with Cauchy and Poisson on methods of solving differential equations in hydrodynamics (1828, 8) suggests that he had read the paper (Fourier 1818) on precisely this matter, which had been published in a Paris journal (Grattan-Guinness 1990, 683–686).

How did Green gain access to these works? While British texts would have been available in the local library, access there to foreign literature is much less certain, even presuming that he had funds available to buy it. The point is particularly perplexing for journals—in particular, the Paris *Mémoires* with its Poisson papers. How did Green *know* that those papers were published there in the first place? Although presentations to the Academy were reported in Paris journals and sometimes abroad, the news did not circulate very much, and Poisson had not given any warning in earlier papers that research in magnetism was in progress. The best chance was that some summary version was translated into a foreign language such as English—and indeed this did happen to summaries of these two papers, in the *Quarterly Journal of Science* (Poisson 1824, 1825).

Each summary paper began with a virtually verbatim repeat of parts of the opening preamble of the parent paper, and then summarised some later results and features. The accounts concentrated mostly on physical and experimental aspects; mathematical procedures were only mentioned (and three formulae quoted in the second summary), although not in a manner to reveal any major novelties. In particular, the divergence theorem (4) was described only in general terms, and with reference to simplification: 'by means of certain transformations, the triple integrals which they contain are reduced to double integrals, and the equations become much more simple' (Poisson 1824, 327). No reader of the time could have guessed that surface integrals were involved; but Green might have been alerted to watch out for the full versions of the papers.

Regarding timetable, the volume of the Paris Academy *Mémoires* containing these papers appeared right at the end of 1826 (Academy of Sciences 1918, 473). Allowing for the usual delay for ships to deliver copies across the channel, one can guess that the spring of 1827 was in hand before Green read at least Poisson's first paper and had his inspiration. Since his essay was to appear in April 1828, this would have given him a maximum of around a year to carry out the research—not an excessive time, even for a part-time mathematician. His motivation was high, most of the required skills and familiarity with the literature were already available—and above all his ideas were fruitful, so that the fruit would grow freely and quickly.

6. OPTIONS FOR PUBLICATION. However, in contrast to this splendid piece of research and development, Green's sales and marketing were hopeless. He cannot be blamed for his scientific isolation in Nottingham, but he was somewhat naive in resorting to the traditions of publication by public subscription. For the increase in scientific activity in Britain in recent years, together with advances in printing technology, had raised the chances and opportunities for publication, especially for an author like him with financial means available to assist with the costs of production. He could have tried Deighton's of Cambridge, who were then publishing quite a lot in mathematics (Grattan-Guinness 1985) and in fact stocked the essay when it came out; or maybe Taylor (now Taylor and Francis), regular

producers of scientific books. He could have written a paper summarising his findings for their *Philosophical magazine*, which was widely distributed in the scientific world: although it did not publish mathematics frequently, there were papers from time to time, and indeed there had been an exchange in there in 1826 on another aspect of potential theory (namely, properties of equipotential surfaces) between Poisson and the Scottish born mathematician James Ivory (Grattan-Guinness 1990, 1190–1195). In fact, if he had felt it proper so to act he could have sought advice from Ivory, the mathematician most conversant with potential theory in Britain at that time. He might also have treated his manuscript as a long paper instead of a short book, and tried to submit it to the Royal Society, or the Cambridge Philosophical Society, or the Royal Society of Edinburgh.

I have no doubt that Green never considered any of these possibilities. His essay went to his 52 supporting subscribers, most of whom could not have read a page of it (Green (H.) 1946, 45–48); and so it vanished from sight. Very rarely has it appeared even in booksellers' catalogues.

7. ON GREEN'S SECOND PERIOD. Green's later career was somewhat less unorthodox than previously, in as much as he was resident at Gonville and Caius College Cambridge (Bromhead's alma mater) from 1833 to 1837 and for some months of 1839–1840 as a Fellow. He had a small overlap in residence with someone capable of understanding his work, indeed the first mathematician to cite the essay; but this was the eccentric Robert Murphy (1806–1843), who spoilt a promising career by financial incompetence.¹

Green published eight papers (and a supplement to one of them), mostly in the *Transactions* of the Cambridge Philosophical Society with Bromhead as communicator. However, his marketing skills were again to the fore: he cited his essay only twice (1871, 120, 192), and on neither occasion did he even give the reader the publication details, never mind a comment to explain its importance.

The other papers fall into two partly related groups, both showing strong French influence in both content and methodology (Burkhardt 1908, ch. 13). One group deals with elastic bodies, which could be construed to be physically bending objects, or else the elastic aether (and perhaps with luck, both at once). The task was to study the propagation of longitude and transverse vibrations; Green also tackled the difficult question of behaviour at the interface between different substances. He sought generality by making no stipulations about the constituted properties of the substances. The principal influence seems to have been the non-molecular studies of elasticity made from 1827 onwards by Cauchy, which had been partly inspired by Fresnel's work in waval optics (Whittaker 1951, ch. 5).

The other group examined the potentials of fluids, which again might cover sound and water, but also the supposed electric and magnetic fluids. Green made this analogy quite explicit in the title of the first of these papers, when referring to the 'laws of the equilibrium of fluids analogous to the electric fluid' (1871, 117).

For methods Green used both his function and theorem, and some of the special results from his essay. He played a little more with operator methods, and produced some solutions in terms of elliptic integrals (although he seemed to be

¹On Green's and Murphy's work see (Cross 1985); the reference to Green's book is in (Murphy 1833, 587). The nature of Murphy's misdemeanours has not been clear; my information comes from a letter of perhaps 1835 to Babbage written by Augustus De Morgan, who was Professor in London University, where Murphy was then trying to make a living (British Library, Additional Mss. 37189, no. 241). Compare (Cannell 1993, 112–113).

unaware of the recently introduced elliptic functions). A paper on the motion of waves and canals took a step towards the approximating asymptotic solution method now known as ‘WKB’ (Schlissel 1977, 309–314), although he limited himself to working within the linearising models of his time. He worked with potentials to the inverse n th power; and on one occasion he required of his potential function that it be invariant under infinitely small rotations, a step that Sophus Lie was to bring (independently) to great generality and prominence 60 years later.

Somewhat separate from Green’s other papers was one dealing with the motion of the ‘simple’ pendulum. This was a favourite topic at this time, a typical example of small-effect science; for the pendulum was required to work to a great degree of accuracy for the purposes of geodesy. Laplace and Poisson, and also F. W. Bessel and G. B. Airy, had been among its many earlier students (Wolf 1889–1891).

8. RECOGNITION. Green’s marketing skills increased at least to the extent that he sent some of these papers to Carl Jacobi (Cannell 1993, 104; the copies are in private possession), and presumably while at Cambridge he gave copies of his essay to William Hopkins, who passed either two or three copies on to the young William Thomson (1824–1907) in 1845.² Then, as is famously known, the essay found its first enthusiastic reader, four years after the death of its author in 1841. Thomson introduced the name ‘Green’s theorem’, and soon came to his ‘method of images’ as result of reading the analysis in the essay of the effect on the electrical charge in a body at an interior/exterior point of a source at a given exterior/interior point (Green 1828, 50–54). He soon arranged for the essay to be reprinted in Crelle’s journal, although it did not appear until 1850–1854.³ Later he and P. G. Tait called the Dirichlet principle ‘Green’s problem’ (Thomson and Tait 1883, arts. 499–518). The name ‘Green’s function’ for functions satisfying conditions like Green’s own is due to Bernhard Riemann and Carl Neumann (Burkhardt and Mayer 1900, art. 18).

Today, Green’s function and his theorem are extolled because of the roles which they continue to play in modern physics and in engineering; but it would be a misunderstanding of history to think that their importance is *due* to these applications. On the contrary, their rise occurred during the period of *classical* physics, when there appeared a mountainous production of books and papers on potential theory and its use in mathematical physics (Bacharach 1883); all the applications mentioned above were involved, and in due course new ones such as thermodynamics and meteorology, and also mathematical economics. All major applied mathematicians took part, along with many minor ones, and some pure mathematicians also (in particular Karl Weierstrass, who sabotaged standard methods of manipulation in 1870 with his famous counter-example to the Dirichlet principle using the inverse tangent function).

Not only Green’s insights and results were used by his successors; his own work, especially the essay, were made available *four times* in the last thirty years of the

²One of these copies of Green’s essay is now kept at Nottingham (Cannell 1993, 105), and another is in Keele University Library; I do not know the location of the third one (if there was one).

³The circumstances of this reprinting of the essay are strange. Firstly, Thomson asked Crelle and not his friend Joseph Liouville, who also edited a journal (still often known after Liouville) and was actively interested in potential theory. Secondly, while Crelle had agreed enthusiastically to the suggestion by 1846 (Green (H.), 41–43), he did not reprint it for several years, and then in three parts over five years.

19th century to an extent surpassing all other literature of his own time. The edition of his works by N. Ferrers appeared in London in 1871, and was reprinted in facsimile in 1903 in (of all places) Paris. The essay itself was also reprinted in facsimile, in 1890 in Berlin, in a series of classic reprints of science; five years later it appeared in an annotated German translation by A. van Oettingen and A. Wangerin, in Wilhelm Ostwald's famous booklet series of editions of major scientific works. Green's successors in the classical phase not only absorbed his contributions into their own heritage; they wanted to read the words of the master himself.

Their modern successors have maintained the tradition; for Ferrers's edition appeared again in 1970, and the *Essay* itself in 1993, in the university of his home town Nottingham, as part of their bicentennial celebrations of their remarkable citizen.

REFERENCES

- 'MAS' denotes the *Mémoires* of the Paris Academy of Sciences, both under that title after the Restoration of 1816 and its Imperial name as the mathematical and physical class of the Institute of France.
- Academy of Sciences 1918. *Procès-verbaux des séances de l'Académie des Sciences tenues depuis la fondation [in 1795] jusqu'au mois d'août, 1835*, vol. 8, Hendaye (Observatoire).
- Bacharach, M. 1883. *Abriss der Geschichte der Potentialtheorie*, Würzburg (Thein).
- Burkhardt, H. 1908. 'Entwicklungen nach oscillirenden Functionen und Integration der Differentialgleichungen der mathematischen Physik', *Jahresbericht der Deutschen Mathematiker-Vereinigung*, 10, pt. 2, xii + 1804 pp.
- Burkhardt, H. and Mayer, F. W. F. 1900. 'Potentialtheorie', in *Encyklopädie der mathematischen Wissenschaften*, vol. 2, pt. A, 464–503 (article IIA7b).
- Cannell, D. M., 1993. *George Green ...*, London (Athlone Press).
- Cross, J. J. 1985. 'Integral theorems in Cambridge mathematical physics, 1830–55', in (Harman 1985), 112–148.
- Fourier, J. B. J. 1818. 'Note relative aux vibrations des surfaces élastiques ...', *Bulletin des sciences, par la Société Philomatique de Paris*, 129–136. [Also in *Oeuvres*, vol. 2, 255–265.]
- Grattan-Guinness, I. 1985. 'Mathematics and mathematical physics at Cambridge, 1815–40 ...', in (Harman 1985), 84–111.
- Grattan-Guinness, I. 1990. *Convolutions in French Mathematics, 1800–1840. From the Calculus and Mechanics to Mathematical Analysis and Mathematical physics*, 3 vols., Basel (Birkhäuser) and Berlin (Deutscher Verlag der Wissenschaften).
- Green, G. 1828. *An Essay on the Mathematical Analysis of Electricity and Magnetism*, Nottingham (the author). [Repr. in *Journal für die reine und angewandte Mathematik*, 39 (1850), 75–89, 44 (1852), 356–374, 47 (1854), 161–211; 1890, Berlin (Mayer and Müller); 1958, Goteborg (Ekelöf); 1993, Nottingham (The University). Also in (Green 1871), 1–115. German trans.: 1895, Leipzig (Engelsmann: Ostwald's Klassiker der exakten Wissenschaften, no. 61).]
- Green, G. 1871. *Mathematical Papers* (ed. N. M. Ferrers), London (Macmillan). [Repr. 1903, Paris (Hermann); 1970, New York (Chelsea).]
- Green, H. 1946. 'A biography of George Green ...', in A. Montagu (ed.), *Studies and Essays in ... Honor of George Sarton*, New York, (Schuman), 545–594.
- Harman, P. M. 1985. (Ed.), *Wranglers and Physicists ...*, Manchester (UP).
- Murphy, R. 1833. 'On the inverse method of definite integrals, ...', *Transactions of the Cambridge Philosophical Society*, 4, 353–408.
- Poisson, S.-D. 1812. 'Mémoire sur la distribution de l'électricité à la surface des corps conducteurs', *MAS*, (1811), pt. 1, 1–92.
- Poisson, S.-D. 1813. 'Remarques sur une l'équation qui se présente dans la théorie des attractions des sphéroïdes', *Nouveau bulletin des sciences, par la Société Philomatique de Paris*, 3 (1812–13), 388–392.
- Poisson, S.-D. 1814. 'Second mémoire sur la distribution de l'électricité à la surface des corps conducteurs', *MAS*, (1811), pt. 2, 163–274.

- Poisson, S.-D. 1824. [Translation of a summary of Poisson 1826a], *Quarterly journal of science*, 17, 317–324. [Original in *Annales de chimie et de physique*, (2)25 (1824), 113–137, 221–223.]
- Poisson, S.-D. 1825. [Translation of a summary of Poisson 1826b], *Quarterly journal of science*, 19, 122–131. [Original in *Annales de chimie et de physique*, (2)28 (1825), 5–18.]
- Poisson, S.-D. 1826a. ‘Mémoire sur la théorie du magnétisme’, *MAS*, 5 (1821–22), 247–338.
- Poisson, S.-D. 1826b. ‘Second mémoire sur la théorie du magnétisme’, *MAS*, 5 (1821–22), 488–533.
- Poisson, S.-D. 1827. ‘Mémoire sur la théorie du magnétisme en mouvement’, *MAS*, 6 (1823), 441–570.
- Schlissel, A. I. 1977. ‘The development of asymptotic solutions to ordinary differential equations, 1817–1920’, *Archive for History of Exact Sciences*, 16, 307–378.
- Thomson, W. and Tait, P. G. 1883. *Treatise on Natural Philosophy*, 2 pts. Cambridge (UP).
- Todhunter, I. 1873. *A History of the Mathematical Theories of Attraction and Figure of the Earth . . .*, 2 vols., London (Macmillan). [Repr. 1962, New York (Dover)].
- Whitrow, G. J. 1984. ‘George Green (1793–1841): a pioneer of modern mathematical physics and its methodology’, *Annali dell’Istituto di Storia della Scienza di Firenze*, no. 2, 47–68.
- Whittaker, E. T. 1951. *History of the theories of aether and electricity. The classical theories*, London (Nelson).
- Wolf, C. 1889–1891. (Ed.) *Mémoires sur le pendule . . .*, 2 pts., Paris (Gauthier-Villars).

School of Mathematics and Statistics
Middlesex University
Enfield, Middlesex
EN3 4SF, England
ivor2@uk.ac.mdx

Lifting Weights

I once, long ago, went into my favorite dean's office to argue once more about good teaching. I remarked that I was teaching a weight lifting class (which he knew I was not). I said that graduation required lifting 250 pounds, that many students got discouraged and dropped out, some repeated the course, and very few graduated. “But last night,” I said, “I had the idea that many more would graduate if I cut the weights in half and graduation would then require lifting one set of 125 pounds, setting them down, and then lifting the second set, thus lifting the 250 pounds.”

Is it reasonable to compare physical muscles to mental muscles? What indeed are good lecturing and good teaching? To what extent does making the teaching clearer and learning easier cut the weights in half? Surely, in the long run and for most courses, developing the student's abilities (mental muscles) to learn is more important than the course content. How often do we invert them?

R. W. Hamming
Naval Postgraduate School
Moterey, CA 93943

Communicating Mathematics: Useful Ideas from Computer Science

Charles Wells

1. INTRODUCTION

1.1. Purpose. This article describes certain ideas originating in the theory and practice of computer science, and shows how the teaching and exposition of mathematics could benefit if these ideas were widely understood by mathematicians and used in their teaching and writing.

These ideas are discussed here because I believe they are important for mathematicians to understand. Some of them are based on theoretical work by computer scientists and others are based on the practice of computer professionals inside and outside academia. Computer scientists would not regard the various concepts as of equal importance, and the whole collection of ideas is nothing like a fair presentation of the current state of computing.

2. SPECIFICATION

2.1. External behavior. A programmer writing a large program may have a tentative conception of how to write the program in terms of subprograms that perform specific tasks. For example, a program for factoring large integers might use a function `PrimeQ:Z → {True, False}` with the property that `PrimeQ [n]` returns `True` if the integer n is prime and `False` otherwise. This description gives the function's *external behavior*¹ but says nothing about how that behavior is implemented. Perhaps the first implementation of `PrimeQ [n]` will test whether any integer k for which $1 < k < \sqrt{n}$ divides n . This might be enough for debugging the program that uses `PrimeQ`, but not fast enough for practical use. Later, an implementation using modern fast techniques could be substituted. Since the *external behavior* of the function `PrimeQ` is the same in either implementation, substituting the new implementation for the old should not introduce new bugs in the program.

In this section, I discuss some issues involved in presenting mathematics to students that can be clarified by this idea of specifying external behavior.

2.2. Is everything a set? One concern of those who study the foundations of mathematics is to show how to develop the main body of modern mathematics from a small number of principles or concepts that are as clear and primitive as possible. The bulk of the work in this direction has been to reduce all mathematical constructions to the primitive notion of “set” and “element of a set” and to

¹Many practicing programmers call this its “functional behavior” but I would avoid that phrase in teaching mathematics students because of confusion with the function concept. My thanks to the referee for suggesting the name “external behavior”.

impose a small number of clear axioms on these primitives.² In carrying this program out, an ordered pair $\langle a, b \rangle$ is typically interpreted as the set $\{\{a, b\}, \{b\}\}$, and a function as a set of ordered pairs with the functional property.

Many mathematicians have taken this approach to mean that every mathematical object is *really* a complicated set. At least, they say this when the topic of foundations comes up. They often don't behave as if they actually believe every mathematical object is a set. Wouldn't you expect a mathematician to be confused, at least momentarily, if you asked which points in the plane had nonempty intersection with the point $\langle 3, 2 \rangle$?³ I maintain that *in practice* many mathematicians regard points as one type of mathematical object, sets as another, and perhaps functions as a third. Since intersection is an operation defined on sets and points are not sets, the question about which points have nonempty intersection with $\langle 3, 2 \rangle$ is to be rejected as meaningless.

The best way to think of the reduction to sets that has been carried out by those who study foundations is that it is a *representation* of mathematics which is desirable for various reasons, for example showing consistency. Although an ordered pair is not a set, it can be represented as a set and that representation may be useful for certain purposes.

2.3. Specification in exposition. My thesis in this section is that we should borrow the idea of specifying external behavior and use *specifications* rather than *definitions* of many common mathematical objects in a way that will exhibit how the objects relate to other types of objects. The formal definition of a concept may require the mention of details of representations used for other purposes (such as consistency proofs) that obscure the way the concept is used in practice. In courses for undergraduates other than in foundations, we should not even attempt to say what sets, pairs, functions and other basic mathematical objects “really are”. What matters is how they relate to other objects.

Recommendation. *In elementary exposition, explain a basic concept by giving a specification of the concept—a carefully written description of the interaction of the object with other mathematical objects.*⁴

Here are two examples based on my text [Wells, 1993], which is aimed at students who have had calculus but no course in abstract mathematics. In these specifications, I use the word “object” to denote any sort of mathematical entity.

Ordered pairs: An *ordered pair* is a mathematical object which is distinct from but completely determined by objects called its *first coordinate* and its *second coordinate*. The ordered pair with first coordinate x and second coordinate y is denoted by $\langle x, y \rangle$.

²Foundations can also be done using category theory [McLarty, 1993].

³This is discussed from a different point of view in [Barr, 1993].

⁴The word “specification” in computer science generally means a description in a formal language of the external behavior of a program, suitable for being transformed by strict rules into an actual program. In this document, the analogy is more with the informal descriptions practicing programmers give of the external behavior of a program, as described in Section 2.1.

It follows that ordered pairs are the same if and only if their coordinates are the same, that is,

$$(\langle x, y \rangle = \langle x', y' \rangle) \Leftrightarrow (x = x' \text{ and } y = y').$$

Thus we have a *method of proof*: To prove two ordered pairs $\langle x, y \rangle$ and $\langle x', y' \rangle$ are the same, prove that $x = x'$ and $y = y'$.

Functions: A function F is a mathematical object which determines and is completely determined by the following data:

- F.1 F has a *domain*, which is a set and is denoted by $\text{dom } F$.
- F.2 F has a *codomain*, which is also a set and is denoted by $\text{cod } F$.
- F.3 For each element $x \in \text{dom } F$, F has a *value* at x . This value is completely determined by x and F and must be an element of $\text{cod } F$. The value of F at x is denoted by $F(x)$.

I am sure these specifications could be improved in various ways and would welcome suggestions concerning them. There may be a better name than “specification” for the practice, too, but it seems clear that the practice should have an explicit name to signal its logical status.

3. SYNTAX AND SEMANTICS. There is a sense in which $2/(4 + 3)$ is $2/7$ and another sense in which $2/(4 + 3)$ is not $2/7$. The *number* $2/(4 + 3)$ is indeed the same number as $2/7$. The *expression* $2/(4 + 3)$ is not the same as the expression $2/7$. For one thing, the expression $2/(4 + 3)$ has seven symbols and the expression $2/7$ has only three.

Syntax is the study of expressions in linguistics or computer science, and *semantics* is the study of how meaning is assigned to expressions. There are two different points to make concerning syntax and semantics: (a) Expressions represent mathematical objects but are not the objects themselves, and (b) expressions have structure.

3.1. Expressions and their denotations. A mathematical expression denotes a mathematical object. The object it denotes is not the expression—the expression is only a representation of the object. In particular, different expressions may denote the same object.

This point of view, that there is an object independent of the expressions that denote it, is often called “Platonist”⁵. By contrast, some assert that the expressions are merely themselves (“everything is syntax”) and that mathematics consist of manipulating these expressions according to precise rules. Presumably, those who hold that attitude will admit that there is an equivalence relation on expressions which identifies different expressions that name the same object from a Platonist’s point of view. In any case, people who hold these differing points of view generally agree on which statements are theorems.

It is my observation that students and teachers at the college level in the USA don’t communicate well with each other because many teachers talk like Platonists,

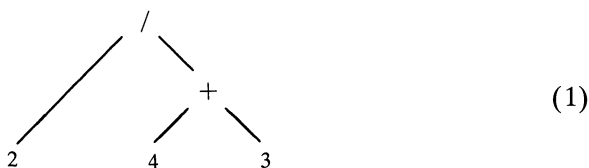
⁵This usage of the word “Platonist” does not imply an endorsement of *all* of Plato’s attitudes toward reality and truth.

but many students have the attitude⁶ that what they need are rules to manipulate the expressions (more about this in 3.3 below). Students who go on to higher mathematics learn to talk as if the mathematical objects were “out there”, but it is noticeable that many college freshmen in calculus courses do not talk that way.

Recommendation. *Teachers and authors of textbooks should make the distinction between syntax and semantics explicit.*

If the student has words for this distinction, he or she may avoid certain types of confusion that result from being unaware of the distinction.

3.2. Parsing. Computer science is intimately concerned with the relationship between syntax and semantics, particularly with *parsing*, which is the explication of the abstract structure of an expression such as $2/(4 + 3)$. This structure is often given as a tree



To *parse* an expression such as $2/(4 + 3)$ is to exhibit its structure. The first task a compiler for a computer language has is to parse the commands of the language, for only then can the commands be executed.

Laborde [1990] notes that many students (these were mostly below the USA freshman level) see expressions such as $2/(4 + 3)$ merely as strings and are not really aware of their abstract structure.

Recommendation. *Introduce informal parsing of mathematical expressions as a learning tool.*

3.3. Mathematics as syntax. Another point of view is that we should *stop* talking like Platonists and go along with the students’ desire for rules for manipulation. Some hold that mathematics is a game of syntax, and that to succeed in mathematics you must master the rules of the game (more about this in Section 4.3). You need not hold that view, however, to realize that a lot of mathematics is accomplished precisely by syntactic manipulation—what else is high school algebra? And even Platonists agree that you have to master the rules of the game.

Recommendation. *Make explicit the allowable syntax for statements about a type of object.*

For example, one can helpfully explain application of functions by adding the following sentence to the specification of function in Section 2.3:

The expression “ $F(x)$ ” is meaningful if and only if “ $x \in \text{dom } F$ ” is true, and in that case “ $F(x) \in \text{cod } F$ ” is true.

⁶Usually, this attitude is unexpressed. I am saying this out of my observations of students rather than what they actually say.

4. FORMAL TRANSFORMATIONS. Formal transformations, called *rewrite rules* in many contexts, are used in many different ways in computing. In general, they work this way: In an expression, you recognize a subexpression as matching the pattern of the left side of a transformation rule, and you rewrite the expression, replacing the subexpression by the right side of the rule. For example, in algebra you may rewrite $a + bx + by$ as $a + b(x + y)$. Computing an integral in freshman calculus can be thought of as recognizing patterns and applying formal transformations, although there may be several possible transformations to apply and the process need not terminate.

Sometimes rewrite rules are applicable to any occurrence of the suitable pattern (they are “context free”) and at other times they depend on specific conditions (“context sensitive”). A notorious example of the latter is L’Hôpital’s Rule. Students resist constraints of this sort [Maurer, 1987].

4.1. Definitions as macros. Most implementations of the C language allow the user to define *macros* that a preprocessor converts into standard C commands. For example, you might want to limit the number of times a program will repeat some action. By writing `#define maxit 20` you defined the macro `maxit` to have the value 20; the preprocessor will replace the word `maxit` with 20 everywhere it appears in the source program. Later, after you have debugged the program, you could change `maxit` to some much larger number and recompile. In general, in contrast to this example, macros can have parameters.

Mathematical definitions play the role of macros in the context of proofs. A colleague of mine in computer science who majored in mathematics as an undergraduate has described how as a student he suddenly caught on that he could do at least B work in most math courses by merely rewriting the definitions of the terms involved in the questions and making a few obvious deductions.

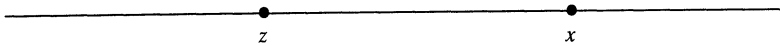
Recommendation. *Encourage students to begin proving a theorem by replacing (some or all of) the words that have definitions with the text of their definitions.*

4.2. Proof by rewriting. Gries [1991], Dijkstra and Scholten [1990] and others have urged that *proofs* be done by applying formal transformations. Many examples may be found in [van Gasteren, 1990] and [Gries and Schneider, 1993]. Proofs are not often done that way by mathematicians except in teaching formal logic. The preferred method is more often the “semantic” approach: the proof proceeds via an understanding of the objects involved rather than by the formal application of rules. Gries, and Dijkstra and Scholten, urge a syntactic approach: express the desired result in a formal language and apply meaning-preserving transformations to it until it becomes a consequence of a known fact. This turns a *proof* into a type of *computation*.

I will now give two proofs of a simple statement about the ordering of the real numbers drawn from Gries [1991]. The statement to prove is

$$(x > z) \Rightarrow ((x > y) \vee (y > z)). \tag{2}$$

4.2.1. Semantic proof. The way I proved (2) when I first saw it was to envision x and z as points on the line placed this way:



There are three different regions into which we can place y . In the right two, $y > z$ and in the left two, $x > y$. End of proof.

This proof is written in English, not in symbolic notation, and it refers to a particular mental representation of the structure in question (the usual ordering of the real numbers).

4.2.2. *Syntactic Proof.* This proof is due to David Gries (private communication). It is based on these principles:

P.1 (Contrapositive) The equivalence of $P \Rightarrow Q$ and $\neg Q \Rightarrow \neg P$.

P.2 (DeMorgan) The equivalence of $\neg(P \vee Q)$ and $\neg P \wedge \neg Q$.

P.3 The equivalence in any totally ordered set of $\neg(x > y)$ and $x \leq y$.

Proof:

$$\begin{aligned}
 (x > z) &\Rightarrow ((x > y) \vee (y > z)) \\
 &\Leftrightarrow \text{by P.1} \\
 \neg((x > y) \vee (y > z)) &\Rightarrow \neg(x > z) \\
 &\Leftrightarrow \text{by P.2} \\
 (\neg(x > y) \wedge \neg(y > z)) &\Rightarrow \neg(x > z) \\
 &\Leftrightarrow \text{by P.3 three times} \\
 ((x \leq y) \wedge (y \leq z)) &\Rightarrow (x \leq z)
 \end{aligned}$$

which is true by the transitive law.

4.2.3. *About the syntactic proof.* There are many advantages to the technique illustrated by the second proof. It holds in any totally ordered set, not just in the real numbers. Each instance of the application of a transformation can be mechanically checked to see that it is correctly applied. (Ingenuity, of course, is still required to *create* the proof.) This mechanical verifiability is certainly not true in the case of the usual mathematical proof; it is notorious that if you read a proof written by someone whose mental representation of the concepts is very different from yours, the proof is next to impossible to follow.

4.2.4. *About the semantic proof.* There are several arguments for the semantic proof. For one thing, many mathematicians prefer the proofs to be written out in English sentences rather than in the symbolic notation of 4.2.2. This view is advocated in the works on mathematical writing by Halmos [1975] (page 42), Steenrod [1975] (page 57), Gillman [1987] (page 15) and Boas [1981]. Another point is that it is easy to make mistakes in checking the application of transformations, particularly when the patterns that must match are complex.

However, the major argument for semantic proofs concerns mental representations.

4.3. Mental representations. Several mathematicians who read the syntactic proof in 4.2.2 in an earlier version of this paper expressed dissatisfaction with it as compared to the pictorial proof preceding it. One objection they gave is that the pictorial proof helps them to *understand* the theorem. I believe that when they say that, they mean they want a *mental representation* of the object involved in the theorem that makes the truth of the theorem obvious or easy to understand.⁷ A

⁷I do not claim that making the truth of the theorem obvious constitutes proof of the theorem. Argumentative philosophers of science who suspect mysticism in this paragraph should observe that I am making a checkable claim about the behavior of mathematicians, not a philosophical claim about Truth.

mental representation of a particular concept is an elaborate *metaphor*. If you can find the right metaphor for a mathematical object, you can follow proofs concerning the object much more easily and you can frequently avoid falling into conceptual traps. (Not only that, it is the mental representation that suggests how the mathematical structure can be used in applications.) Following the proof line by line may convince one that the theorem is correct, but it gives no understanding unless the proof aligns in some sense with one's internal representation of the concept.⁸ Indeed, the hope is that it will *refine* or *reform* one's inner representation of the concept.⁹

The statements in the preceding paragraph about mental representation are controversial: they reflect my own position but not the position of all mathematicians. Those statements caused far more correspondence than any others in this article. Some said that they do not use mental representations. For them, mathematics is all syntax. Others were dismayed by the syntactic proof and felt that the primary purpose in teaching was to transmit to the students useful mental representations of the concepts. Thus there are two kinds of mathematicians: Those for whom the mental representation (they often say "intuition" or "understanding") is paramount, and those who insist that syntax is primary. The gulf between these two kinds of mathematicians is vast. It is as if we were two different kinds of intelligent beings who are deluded into thinking we are communicating with each other. (But we *do* communicate.)

It is not unreasonable to assume that some of our students tend one way and some the other. I am in the mental representation camp, but in recent years, I have used syntax and transformations of statements in class more than I used to, and I believe it makes a difference for the better to the students.

4.4. Explicit use of logic. Even mathematicians in the mental representation camp use computation in proofs. Finding a suitable representation of an object that allows one to compute is as old as mathematics. However, most mathematicians rarely compute explicitly with the rules of logic as exemplified in 4.2.2. I believe that mathematicians should be aware of the possibilities of this approach, in teaching if not in their own research.

Recommendation. *Transmit your mental representation of concepts whenever you can, but also give proofs as explicit logical calculations when appropriate, because that provides the student with a second way to deal with the problem and provides him or her with the tools to carry out similar proofs.*

Related to explicit mention of logical concepts is the idea of giving a rule of inference for particular types of objects. For example, setbuilder notation has the following rules of inference: From the statement $a \in \{x|P(x)\}$ (where $P(x)$ is a predicate) one can deduce $P(a)$, and from the truth of $P(a)$ one can deduce

⁸Some readers commented that after reading the syntactic proof in 4.2.2 they suddenly understood that in a totally ordered set the condition to be proved is merely the contrapositive of transitivity. So a syntactic type of proof can be illuminating, too.

⁹There is a sizeable research literature on the subject of mathematicians' and students' mental representations of mathematics. See the articles in [Schoenfeld, 1987a], particularly [Schoenfeld, 1987b] and [Maurer, 1987], as well as [Harel and Dubinsky, 1992], [Miller, 1987] and the discussion starting in the second column of page 1187 of [Devlin, 1992].

$a \in \{x|P(x)\}$. My students have found this explicit mention of allowable inference to be helpful.

Recommendation. *Give explicit rules of inference for concepts when they are introduced.*

Note that this was done in the specification for ordered pairs in Section 2.3.

Lamport [1993] provides a detailed model for presenting proofs in a structured way that have the potential for clarifying proofs in either style, symbolic or based on mental representations.

5. TYPES AND POLYMORPHISM. In Pascal and in other typed programming languages, if you declare a variable to be of type Boolean and then try to set it equal to 3 the compiler gives you an error message. This is an example of mismatched *types*.

5.1. The multiplicity of types. The further students go in mathematics, the more different types of data they have to deal with. The typical second or third semester calculus course introduces two and three dimensional vectors, matrices, functions $F: \mathbb{R} \times \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ (“scalar fields”), functions $F: \mathbb{R} \rightarrow \mathbb{R} \times \mathbb{R} \times \mathbb{R}$ (“paths in space”) and functions $F: \mathbb{R} \times \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R} \times \mathbb{R} \times \mathbb{R}$ (“vector fields”) to know about, as well as all the objects of single-variable calculus. At some point we cease to be able to distinguish all these different things by different letters and typefaces, and the students have to learn to understand the types of the expressions they see by reading the surrounding text.

More than that, the meanings of the operator symbols in the formulas at that level may depend on the types of the operands. Consider the “ \times ” symbol. In the expression 3×5 it denotes numerical multiplication. If A and B are three dimensional vectors, $A \times B$ denotes the vector product, but if they are sets, it denotes the cartesian product. In computer science terms, the “ \times ” symbol is “polymorphic”, in the sense that its meaning is dependent on the types of its arguments.

Recommendation. *Use the concepts of type and polymorphism explicitly to help students to understand and avoid the traps of type confusion.*

Most students now have had some experience with programming languages that use typing. I have found that to refer to types explicitly is helpful. I write “TYPE ERROR” on their homework when such a mistake is made, and sometimes when I forget to put the little arrow over a symbol for a vector on the blackboard, I get a chorus of “TYPE ERROR” from the class, which I think is great.

5.2. Teaching conceptual distinctions. A particularly bad typing error concerns functions. A function $F: A \rightarrow B$ has a value $F(a)$ at each element of A and, particularly in undergraduate mathematics, it may be given by an expression that is used to compute its values. The function, its value at an input, and an algorithm for computing it are three different mathematical objects that must be kept distinct.¹⁰ Students often learn to cope with this in calculus courses by gaining an

¹⁰Our mathematical ancestors confused these, too [Selden and Selden, 1992], [Sfard, 1992].

implicit understanding of the differences. Because the distinctions are not *explicit*, the students' understanding is not on firm intellectual ground.

For the most part, we do not try very hard to convince our students that a function is not the same thing as its defining expression or its values. I have pushed that point in some courses I teach and it helps. If you really want them to know it, of course, you have to test them on it. Far too many mathematicians are unwilling to try testing first and second year students on conceptual content of this sort because they feel that most students will fail the questions. In fact, students can be taught conceptual distinctions if the teacher starts slowly and asks very simple questions at first. All you have to be willing to do is give up about 20% of the "content" of the course. I believe the gain far outweighs the loss, in courses for non-math-majors and math majors alike.

Recommendation. *Expect conceptual understanding at the appropriate level from all students in any course, and test them on it.*

6. SELF-MONITORING

6.1. Name your behavior. The New Hacker's Dictionary [Raymond, 1991] is a compilation of computer jargon which has to be one of the most enjoyable dictionaries ever composed. One thing that becomes noticeable if you read it straight through is the number of words and phrases hackers¹¹ have invented to describe their own mental states or behavior while working. This is discussed in the introduction to the Dictionary. "Juggling eggs", for example, refers to the necessity of keeping a lot of details in your head while modifying a program—with the consequence that an interruption can cause you to scramble the program (this is a paraphrase of the book's definition).

Of course, that is a phenomenon familiar to research mathematicians. You can't spend short, separated pieces of time trying to understand a complicated mathematical phenomenon; you need the time to concentrate and to get it all in your head at once—to be in "hack mode" (p. 190 of [Raymond, 1991]). The point is, mathematicians don't have a name for this, as far as I know. Computer hackers do. We should emulate them.

Computer people give names to their own counterproductive behavior quite freely—look up "creationism", "kluge", "mung" and "thrash" in [Raymond, 1991]. (I have personally munged several chapters of my class notes and had to tear them up and start over.) It would be particularly helpful to give names to common mistakes made by students in math courses. Pólya [1948] emphasized the importance of introducing notation for the quantities in a problem. It is equally important to name *behaviors*, both useful and harmful, that occur in problem solving.

Recommendation. *Describe and name the common kinds of mistakes students make.*

If there is a memorable name for such mistakes the student will be more likely (and will find it easier) to monitor his or her behavior. Self-monitoring is widely

¹¹A hacker is someone who programs for the sake of programming—although useful tools may result, they are not the primary motivation—and who enjoys learning and using the obscure features and behavior of various operating systems and programming languages. The latter-day meaning of someone who breaks into private systems is not intended here.

cited in the educational literature as one of the properties that distinguish good students from poor ones. See [Resnick, 1987], pages 25–27, and [Schoenfeld, 1987c].

One typical mistake occurs when using concepts that by definition require the existence of something. For example, for integers m and n , m divides n if there is an integer q such that $n = qm$. When faced with proving that if m divides both n and p , then m divides $n + p$, a common mistake is to write down the assumptions as $n = qm$ and $p = qm$, using the same q . Recently in class I exclaimed, “If Bob and Ray are both married, that doesn’t mean they have the same wife!” If I had said this on the network, such behavior might have become known as “existential bigamy” (or some such phrase).

It would be useful to come up with punchy names for good behaviors such as the following:

- Working examples before attacking the general case of a problem.
- Naming all the variables in a problem.
- Checking special cases of a statement to see if it is consistent with the rest of mathematics. (This is sometimes called a “sanity check”).

We should also name destructive behaviors such as these:

- Forgetting to check trivial cases. Hackers have an analogous error they call a “fencepost error”—getting the bounds wrong in a loop is an example.
- Proving an implication backward—in other words, being asked to prove $P \Rightarrow Q$ and coming up with a proof of $Q \Rightarrow P$. This is distressingly common among students whom I teach mathematical reasoning.
- Reading variable names as labels ([Nesher and Kilpatrick, 1990], pages 101–102) so that a statement such as “There are six times as many students as professors” gets translated as $6s = p$ instead of $6p = s$ (where p and s have the obvious meanings).

6.2. Context. People who have grown up together or who have worked in the same place for a long time have what have been called “high-context” conversations with many elliptical references to shared ideas, opinions and experiences. A group of people from different cultures or who live in a large city and don’t know each other well will have “low-context” conversations with more made explicit and more attempt to avoid the assumption that the others share one’s point of view.

Mathematicians seem to avoid connotative, high-context conversation about doing mathematics even though the potential is there for communicating quite complex ideas about the subject and about one’s behavior while doing it. It is clear from the New Hacker’s Dictionary that hackers do have high-context communication about these things. Mathematicians have been trained explicitly and through bitter experience that connotations can mislead when doing a proof. Perhaps many of us have mistakenly applied this lesson to other areas of our life, insisting on low-context conversation even when high-context conversation is possible. Or perhaps bright people who are not particularly talented at picking up social context are attracted to mathematics.

I don’t know how we can change the mathematical culture to encourage high-context interaction. Perhaps the spread of email will help; linguists have discovered that linguistic innovation spreads much more rapidly in a language spoken by a large number of people in contact with each other than it does in small groups.

ACKNOWLEDGMENTS. This work has benefited by discussion with and suggestions and corrections from Atish Bagchi, Michael Barr, Stephen J. Bevan, J. E. Fritz, Leonard Gillman, David Gries, C. A. R. Hoare, Colin McLarty, Eric S. Raymond, Daniel M. Rosenblum, Guy Steele, and Leon Sterling. I am particularly grateful to Eric Raymond for many detailed suggestions and insights, particularly in Sections 2 and 6, and for organizing an electronic mailing list for discussions of earlier drafts of this article and of [Bagchi and Wells, 1993].

REFERENCES

- [Bagchi and Wells, 1993] Atish Bagchi and Charles Wells. *The varieties of mathematical prose*. Available by anonymous FTP from ftp.cwru.edu. It is the file mathrite.dvi in the directory math / wells, 1993.
- [Barr, 1993] Michael Barr. *Functional set theory*. Preprint, Department of Mathematics, Burnside Hall, McGill University, 805 Sherbrooke St. West, Montréal, P.Q., Canada H3A 2K6. It is available by anonymous FTP from triples.math.mcgill.ca, in the files variable.sets.tex.2 and variable.sets.dvi.2 in the directory pub / barr, 1993.
- [Boas, 1981] Boas. *Can we make mathematics intelligible?* American Mathematical Monthly, 88(10):727–731, December 1981.
- [Devlin, 1992] Keith Devlin. *Computers and mathematics*. Notices of the American Mathematical Society, 39(10):1186–1188, 1992.
- [Dijkstra and Scholten, 1990] E. W. Dijkstra and C. S. Scholten. *Predicate Calculus and Program Semantics*. Springer-Verlag, 1990.
- [Gillman, 1987] Leonard Gillman. *Writing Mathematics Well*. Mathematical Association of America, 1987.
- [Gries and Schneider, 1993] David Gries and F. B. Schneider. *A Logical Approach to Discrete Mathematics*. Springer-Verlag, 1993.
- [Gries, 1991] David Gries. *Teaching calculation and discrimination: A more effective curriculum*. Communications of the ACM, 34:44–55, 1991.
- [Harel and Dubinsky, 1992] Guershon Harel and Ed Dubinsky, editors. *The Concept of Function*, volume 25 of *MAA Notes*. Mathematical Association of America, 1992.
- [Kieran, 1990] Carolyn Kieran. *Cognitive processes involved in learning school algebra*. In *Mathematics and Cognition*, Pearla Nesher and Jeremy Kilpatrick, editors, ICMI Study Series, pages 96–112. Cambridge University Press, 1990.
- [Knuth et al., 1989] Donald E. Knuth, Tracy Larrabee, and Paul M. Roberts. *Mathematical Writing*, volume 14 of *MAA Notes*. Mathematical Association of America, 1989.
- [Laborde, 1990] Colette Laborde. *Language and mathematics*. In *Mathematics and Cognition*, Pearla Nesher and Jeremy Kilpatrick editors, ICMI Study Series, pages 53–69. Cambridge University Press, 1990.
- [Lampert, 1993] Leslie Lamport. *How to write a proof*. Technical Report SRC-094, Digital Systems Research Center, February 1993. Available by FTP from gatekeeper.dec.com in the directory / archive / pub / DEC / SRC / research-reports. It is file SRC-094.ps.2.
- [Maurer, 1987] Stephen B. Maurer. *New knowledge about errors and new views about learners: What they mean to educators and more educators would like to know*. In *Cognitive Science and Mathematics Education*, Alan Schoenfeld, editor, pages 165–188. Lawrence Erlbaum Associates, 1987.
- [McLarty, 1993] Colin McLarty. *Numbers can be just what they have to*. Nous, 27:487–498, 1993.
- [Miller, 1987] Arthur I. Miller. *Imagery in Scientific Thought*. MIT Press, 1987.
- [Nesher and Kilpatrick, 1990] Pearla Nesher and Jeremy Kilpatrick. *Mathematics and Cognition*. ICMI Study Series, Cambridge University Press, 1990.
- [Pólya, 1948] G. Pólya. *How to Solve It*. Princeton University Press, 1948.
- [Raymond, 1991] Eric Raymond. *The New Hacker's Dictionary*. The MIT Press, 1991.
- [Resnick, 1987] Lauren B. Resnick. *Education and Learning to Think*. National Academy Press, 1987.
- [Schoenfeld, 1985] Alan Schoenfeld. *Mathematical Problem Solving*. Academic Press, 1985.
- [Schoenfeld, 1987a] Alan Schoenfeld, editor. *Cognitive Science and Mathematics Education*. Lawrence Erlbaum Associates, 1987.
- [Schoenfeld, 1987b] Alan Schoenfeld. *Cognitive science and mathematics education: An overview*. In *Cognitive Science and Mathematics Education*, Alan Schoenfeld, editor, pages 1–32. Lawrence Erlbaum Associates, 1987.
- [Schoenfeld, 1987c] Alan Schoenfeld. *What's all the fuss about metacognition?* In *Cognitive Science and Mathematics Education*, Alan Schoenfeld, editor. Lawrence Erlbaum Associates, 1987.

- [Selden and Selden, 1992] Annie Selden and John Selden. *Research perspectives on conceptions of functions*. In *The Concept of Function*, Guershon Harel and Ed Dubinsky, editors, volume 25 of *MAA Notes*, pages 1–16. Mathematical Association of America, 1992.
- [Sfard, 1992] Anna Sfard. *Operational origins of mathematical objects and the quandary of reification—the case of function*. In *The Concept of Function*, Guershon Harel and Ed Dubinsky, editors, volume 25 of *MAA Notes*, pages 59–84. Mathematical Association of America, 1992.
- [Steenrod *et al.*, 1975] Norman E. Steenrod, Paul R. Halmos, Menahem M. Schiffer, and Jean A. Dieudonné. *How to Write Mathematics*. American Mathematical Society, 1975.
- [van Gasteren, 1990] A. J. M. van Gasteren. *On the Shape of Mathematical Arguments*, volume 445 of *Lecture Notes in Computer Science*. Springer-Verlag, 1990.
- [Wells, 1993] Charles Wells. *Discrete mathematics*. Class notes, Department of Mathematics, Case Western Reserve University, 1993.

Department of Mathematics
Case Western Reserve University
10900 Euclid Ave.
Cleveland, OH 44106-7058
cfw2@po.cwru.edu

PICTURE PUZZLE
(from the collection of Paul Halmos)



Setting out on his career: Who was he?
(see page 426.)

Order and Chaos on Your Desk

Susan Bassein

This paper describes a physical system which is easy to build—and can fit in a small clearing on your desk—and whose dynamics can be easily varied to demonstrate some fundamental concepts of dynamical systems. And, because its dynamics take place in one dimension, they are simple to analyze: although the restriction to one dimension excludes some of the phenomena which are responsible for much of the current interest in dynamical systems [1, 4], it allows one to draw pictures which illuminate some of the central ideas of the subject [1, 2, 3]. On the other hand, examples of chaotic systems (e.g., [1, 4]) are typically either difficult to realize in practice or result in dynamics in dimension 2 or higher. For example, the complicated (albeit fascinating!) dynamics of a periodically, externally forced, damped pendulum in $S^1 \times \mathbb{R}$ are described in [1, 4].

To understand what the challenge of designing a system with one-dimensional dynamics entails, let us see why the dynamics of the forced pendulum take place in $S^1 \times \mathbb{R}$. As shown in Figure 1, the state of the pendulum system can be described by three parameters: the position θ of the pendulum (in S^1), the (angular) velocity $d\theta/dt$ of the pendulum (in \mathbb{R}), and the phase of the oscillation of the external force applied at that time (in S^1). If we take periodic “snapshots” of the system at the moments that the oscillation of the external force passes through some given, fixed phase, then the state of the system at those moments can be described by the remaining two parameters. Since the state in the next snapshot is a function of the state in the current snapshot, we obtain a function from $S^1 \times \mathbb{R}$ to $S^1 \times \mathbb{R}$ whose iteration describes the sequence of states observable in the sequence of snapshots. For a system’s dynamics to take place in \mathbb{R} instead, its state in each snapshot must be described completely by a single parameter.

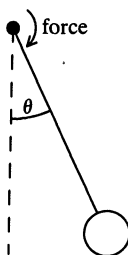


Figure 1. A forced pendulum

My original design was a water tank regulator (which I never built), but because electricity is a much more tractable fluid in practice (and electrons are much less of a nuisance when they slosh all over your desk), I replaced the flow of water with the flow of electrons. An intermediate design, which I built with the help of engineers Chuck Iverson and Skip Korhonen, combines electronic and mechanical

components and can be made to produce either a rhythmic or delightfully chaotic noise. But the version that works most reliably and is simplest to build, control, and model is an electronic approximate simulation of those other designs. The electro-mechanical and electronic systems are pictured in Figure 2.

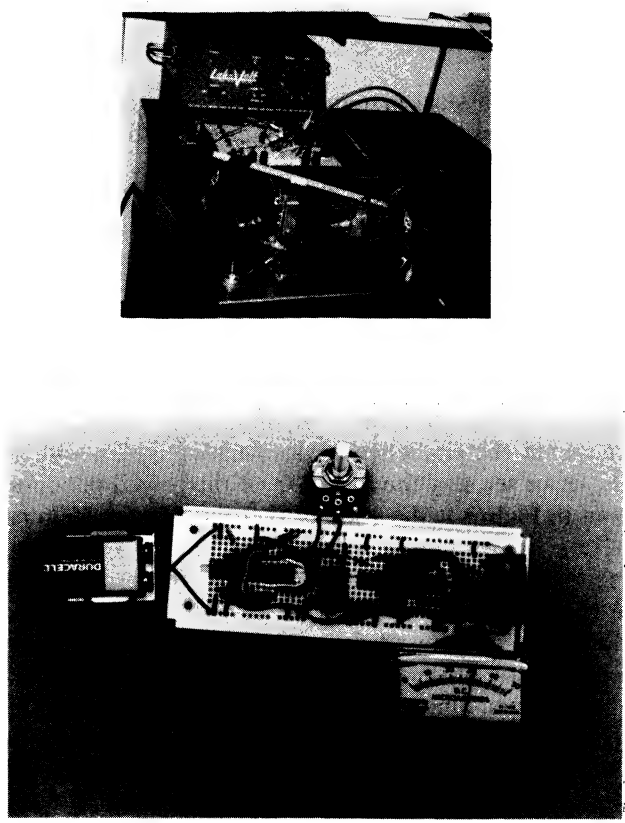


Figure 2. The electro-mechanical and electronic systems

While the physical implementation in each of the three designs differs, the basic components, which are listed in the following table, and the organization of each system, which is illustrated in Figure 3, are the same; the precise specification of the electronic design appears in the Appendix.

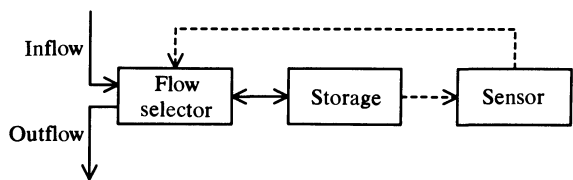


Figure 3. Basic system design

Component	Water flow	Electro-mechanical	Electronic
Storage	Water tank	Capacitor	Capacitor
Inflow	Pipe	+9 volt supply	+9 volt supply
Outflow	Pipe	Ground	Ground
Sensor	Nozzle and mechanical arm	Solenoid and mechanical arm	Capacitor and voltage comparator
Flow selector	Valves	Mechanical switches	Electronic switches

The operation of the system, illustrated in Figure 4 for the electronic design, is as follows. Let $x = x(t)$ be the quantity (of water or charge) in the storage device at time t . The state of the sensor at time t is described by a single, non-negative real number parameter $s = s(t)$ (the angular displacement of a mechanical arm or the charge stored in an auxiliary capacitor). In its rest state $s = s_0$ at some time t_0 , the sensor measures $x(t_0)$ and then s rises to a maximum value which depends on $x(t_0)$ and then decreases to return to s_0 at a time $t = t_1$, which also depends on $x(t_0)$. There is a threshold value s_d of s such that when $s \leq s_d$, the storage device is connected to the inflow and when $s > s_d$, the storage device discharges through the outflow.

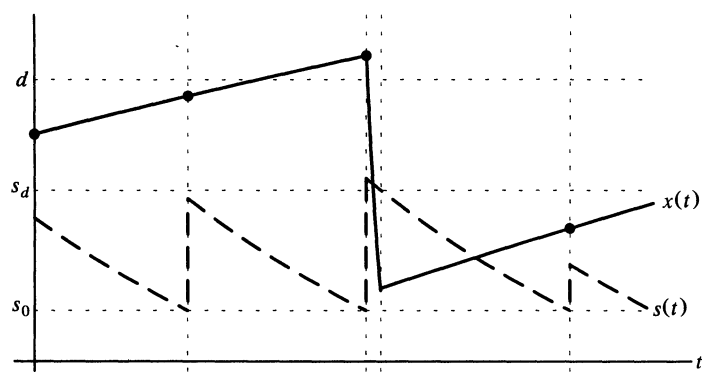


Figure 4. Operation of the sensor in the electronic system

The net result of this arrangement is that if $x(t_0)$ is below some threshold value d (which corresponds to s_d), then the voltage $x(t_1)$ at the next sensing will be greater than $x(t_0)$, but the more $x(t_0)$ is above d , the greater will be the amount that $x(t_1)$ will be less than $x(t_0)$. Let F be the function which gives $x(t_1) = F(x(t_0))$; in the Appendix we show that for the electronic design, if R is the outflow resistance to ground, then F is given approximately by the formula

$$F(x) = \begin{cases} 9 - (9 - x) \left(\frac{4.03}{x + 2.11} \right)^{0.392} & \text{if } 2.11 \leq x \leq 4.35 \\ 9 - \left(9 - x \cdot \left(\frac{6.46}{x + 2.11} \right)^{2 \times 10^5 / R} \right) \left(\frac{4.03}{6.46} \right)^{0.392} & \text{if } 4.35 < x \leq 9 \end{cases}$$

which we will use throughout the remainder of this paper. (If $0 \leq x < 2.11$, then the sensor waits until x rises to 2.11 and then resumes cycling.) Here the discharge threshold is $d = 4.35$. The graph of F with $R = 20,000$ (ohms) appears in Figure 5.

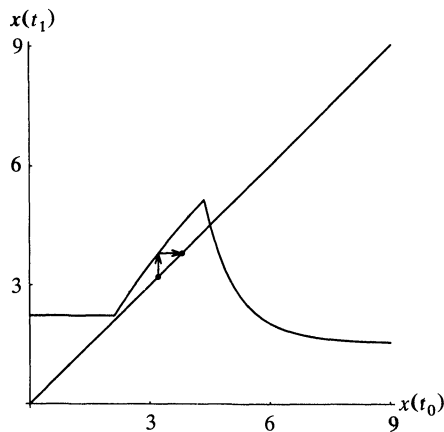


Figure 5. $x(t_1)$ as a function of $x(t_0)$ and the path of x

Since the value of x at the moment s returns to s_0 for a sensing determines the value of x at the next sensing, the sequence of those values of x (in \mathbb{R}) forms the dynamics of the system. (In practice, we use the maximum value reached by the sensor after each sensing to monitor the state of the system.) In terms of F , the succession of those states starting, say, at $t = 0$, will form the sequence of iterates $F^n(x(0))$ for $n \geq 0$, where “ F^n ” means the composition of n copies of F . Figure 5 shows the standard illustration of how each step of the iteration follows the graph of F by moving from a point $(x(t_0), x(t_0))$ through $(x(t_0), x(t_1)) = (x(t_0), F(x(t_0)))$ to $(x(t_1), x(t_1))$.

We can vary the dynamics of the system by controlling the rate of outflow (by adjusting the outflow valve or by varying the resistance R between the storage device and ground): as illustrated in Figure 6 and proved below, a slow outflow makes the sequence of values of x at the moments of sensing approach a fixed point, a moderate outflow produces stable periodic behavior, and a fast outflow can result in chaos. The standard analysis [1, 2, 3] shows that the “attracting fixed point” illustrated in the left picture in Figure 6 results from $|F'(x)| < 1$ at the fixed point; a simple computation shows that this will happen for $R \geq 55,900$. (Note that in practice, resistances can only be specified to an accuracy of 5 or 10%.) The “attracting period 2 orbit” illustrated in the middle picture results from $|F'(x)| > 1$ at the fixed point (which makes it “repelling”) and a pair of points on the graph, positioned symmetrically across the diagonal, at which $|(F^2)'(x)| < 1$; this will happen for $52,400 \leq R \leq 55,800$. Attracting orbits of period 4 (or period 3, or higher periods, if you have a delicate touch on the control) can also be

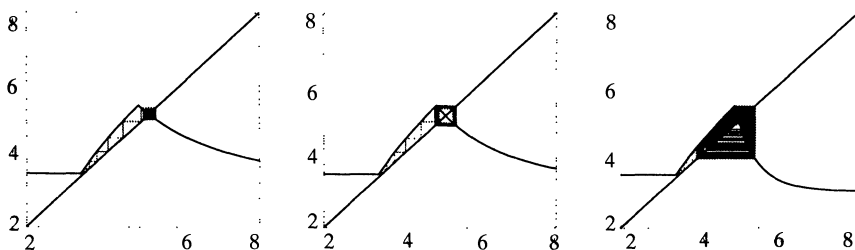


Figure 6. Dynamics for $R = 61,000$, $R = 53,000$, and $R = 20,000$

obtained from the electronic system from still lower values of R . And we show below that for $13,500 \leq R \leq 30,000$, the system exhibits chaotic dynamics on an interval of values for x ; in particular, for this range of values of R , the system does not have any attracting periodic orbits which would make the *observable* dynamics periodic even in the presence of *theoretical* chaos, which would be hidden in a Cantor set of x values. (In fact, it is possible, with a less simplistic analysis than is presented below, to prove that the system will behave chaotically on an even wider interval of R values, but the extra work would not be justified by the accuracy of the model, so we omit it; see [2, 3] for a deeper and more general view of one-dimensional dynamical systems.)

We recall the definition of chaos given in [1] and [3]:

Definition. A map F from a metric space M to itself is *chaotic* on M if

1. F has sensitive dependence on initial conditions: there exists a $\delta > 0$ such that for any $x \in M$ and any neighborhood U of x , there exists a $y \in U$ and an $n > 0$ such that $|F^n(x) - F^n(y)| > \delta$;
2. F is topologically transitive: for any pair U, V of open sets, there is an $n > 0$ such that $F^n(U) \cap V \neq \emptyset$; and
3. Periodic points of F are dense in M .

Notation. Let $d = 4.35$, $a_0 = 2.11$, $b = F(d) = 5.1352\dots$, and $a = F(b)$; note that b is independent of R . Thus, the domain of F is $[a_0, 9] \supset [a_0, b]$. Figure 7 illustrates the relationships between each of these quantities.

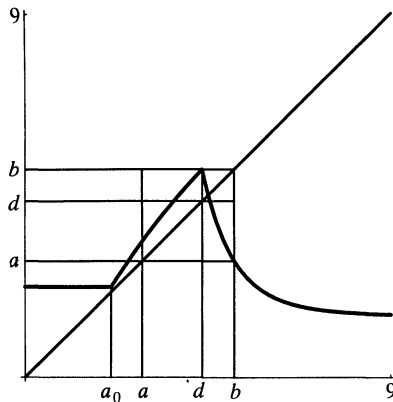


Figure 7. Important points on the graph of F

Remarks. We make a few observations that will be used in the proofs below. First note that for $x \in [a_0, d]$, the value of $F(x)$ does not depend on R and for every fixed $x \in (d, b]$, the value of $F(x)$ is an increasing function of R because $6.46/(x + 2.11)$ is less than 1. We compute $F(a_0) = 2.23330\dots > a_0$. Only slightly tedious derivative computations verify that for all $R > 0$, the function F is increasing on $[a_0, d]$ and concave down on (a_0, d) and decreasing on $[d, b]$ and concave up on (d, b) ; thus b is an absolute maximum for F and $a < b$. Another computation shows that the left derivative of F at $x = d$ is greater than 1.06 (independent of R), hence $F'(x) > 1.06$ on all of (a_0, d) .

Proposition 1. *If $R \geq 11,600$, then for all $x \in [a_0, b]$ there exists an $N \geq 0$ such that $F^n(x) \in [a, b]$ for all $n > N$.*

Proof: A computation shows that the condition $R \geq 11,600$ guarantees that $F(b) > a_0$ so that $F([a_0, b]) \subset [a_0, b]$ and therefore we may iterate F on $[a_0, b]$. First suppose $x \in [a_0, d)$: it follows from the remarks above that $F(x) - x \geq F(a_0) - a_0 > 0.12$ so there is an $N > 0$ such that $F^N(x) \in [d, b]$. If $x \in [d, b]$ instead, let $N = 0$, so in all cases we have $F^N(x) \in [d, b]$. Then, since F is decreasing on $[d, b]$, it follows that $F^{N+1}(x) \in [F(b), F(d)] = [a, b]$. Finally, we have $F([a, b]) \subset [a, b]$: if $x \in [a, d)$ (in the case $a < d$), then $F(x) \in [F(a), F(d)) \subset [a, b]$, else if $x \in [d, b]$, we also have $F(x) \in [a, b]$ by the previous reasoning. QED

Proposition 2. *If $13,500 \leq R \leq 30,000$, then for every non-empty open interval $I \subset [a, b]$, there is an $n \geq 0$ such that $F^n(I) = [a, b]$.*

Proof: The condition $R \leq 30,000$ guarantees that $a < d$ and that $F(a) \leq d$, from which it follows that there are two numbers, $d_1 \in [a, d)$ and $d_2 \in (d, b]$, such that $F(d_1) = F(d_2) = d$. For an interval J , let $|J|$ denote its length. We prove the proposition by showing that if any two of d, d_1 , and d_2 are in I , then $F^2(I) = [a, b]$, otherwise either $|F(I)| > 1.06|I|$ or $|F^2(I)| > 1.008|I|$, so iterates of I will grow in length until the first condition is satisfied.

Suppose two of d, d_1 , and d_2 are in I . Since $d_1 < d < d_2$, if d_1 and d_2 are in I , then so is d , so in any case, $d \in I$. We use the fact that F is increasing on $[a, d]$ and decreasing on $[d, b]$: we have either $I \supset [d_1, d]$ or $I \supset [d, d_2]$; in the first case we have $F(I) \supset [F(d_1), F(d)] = [d, b]$ and in the second we have $F(I) \supset [F(d), F(d_2)] = [d, b]$ also, so $F^2(I) \supset [F(b), F(d)] = [a, b]$, as required.

Now suppose that no two of d, d_1 , or d_2 are in I . If $d \notin I$, then either $I \subset [a, d]$ or $I \subset [d, b]$. In the first case, since $F'(x) > 1.06$ on I , we have $|F(I)| > 1.06|I|$. In the second case, on I we have $|F'(x)| \geq |F'(b)| = |F'(F(d))|$; a derivative computation shows that $F'(F(d))$ as a function of R is decreasing to the left of and increasing to the right of an absolute minimum near $R = 19,747$ and values less than -1.44 at $R = 13,500$ and $30,000$, respectively. Thus, $|F'(x)| > 1.44$ on I and $|F(I)| > 1.44|I| > 1.06|I|$.

Finally, suppose $d \in I$ but neither d_1 nor d_2 is. Let $I = (d - x, d + y)$. Then by the same reasoning as above, F expands the interval $(d - x, d)$ to at least length $1.06x$. We claim that F expands the interval $(d, d + y)$ to at least length $2.07y$. The slope of the line between the points $(d, F(d))$ and $(d + y, F(d + y))$ is more negative than the slope of the line between $(d, F(d))$ and $(b, F(b))$, which is $(F(b) - F(d))/(b - d)$. This expression is an increasing function of R because the only part of it that depends on R is $F(b)$. For $R = 30,000$, that slope is less than -2.07 , which proves the claim. It follows that F multiplies the length of I by at least a factor of

$$\frac{\max\{1.06x, 2.07y\}}{x + y} \geq \frac{1.06 \cdot 2.07}{2.07 + 1.06} > 0.7.$$

Since neither d_1 nor d_2 is in I and $F(d_1) = F(d_2) = d$, it follows that $F(I)$ lies above d , so $F(I) \subset [d, F(d)]$. By the reasoning of the previous paragraph, we have $|F^2(I)| > 1.44|F(I)| > 1.44 \cdot 0.7|I| > 1.008|I|$, as required. QED

Corollary. *If $13,500 \leq R \leq 30,000$, then F is chaotic on $[a, b]$ and has no attracting periodic orbits.*

Proof: To prove that F has sensitive dependence on initial conditions, we let $0 < \delta < (b - a)/2$: if I is an open interval with $x \in I \subset [a, b]$, we have $F^n(I) = [a, b]$ for some $n \geq 0$, so choose a $y \in I$ which F^n maps to whichever of a or b is further from $F^n(x)$. F is topologically transitive because every non-empty open interval eventually maps to all of $[a, b]$. Every non-empty open interval $I \subset [a, b]$ contains a periodic point because if $F^n(I) = [a, b]$, then F^n has a fixed point in I , hence there is a periodic point in I whose period is no more than n . Finally, if there were an attracting periodic orbit, there would be an interval around each of its points which was contracted toward the orbit points by repeated application of F ; Proposition 2 says this cannot happen. QED

APPENDIX: THE ELECTRONIC DESIGN. As illustrated in Figure 8, the electronic design consists of a collection of resistors and capacitors, a potentiometer to control the outflow resistance, a 9 volt battery, and four CMOS integrated circuits: two “op amps” (operational amplifiers, 3140) and two transistor-pair chips (4007) which are wired to make single-pole double-throw switches as described in [5]. Figure 2 showed the arrangement on a breadboard, with a meter (microamperes), an additional op amp (3140) to provide a high-impedance input to the meter, and a 1000 μf capacitor across the battery to protect the circuit when the battery is connected or disconnected. If a breadboard and a high-impedance input meter (or an oscilloscope) can be obtained from an electrical engineering, computer science, or physics department, the rest of the parts (battery not included) should cost at most \$15.

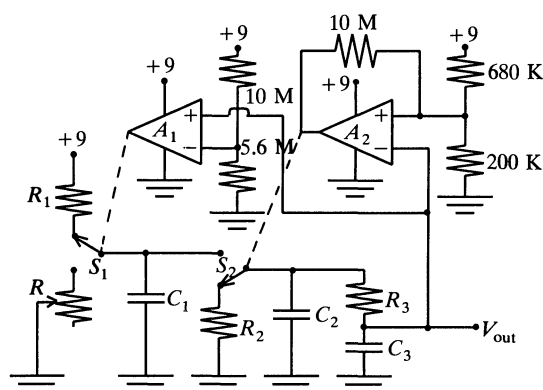


Figure 8. Circuit diagram

$R_1 = 510\text{K}$	$C_1 = 10\mu\text{f}$	$S_1 = S_2 = 4007$
$11\text{K} \leq R \leq 61\text{K}$	$C_2 = 0.1\mu\text{f}$	$A_1 = A_2 = 3140$
$R_2 = 10\text{M}$	$C_3 = 0.1\mu\text{f}$	
$R_3 = 200\text{K}$		

Briefly, the circuit works as follows to produce the behavior illustrated in Figure 4 and the formula given above for F . Recall that if a capacitor with capacitance C and initial voltage V_0 is connected to an external voltage V_e through a resistance R , its voltage V at time t will be given by the RC -formula: $V = V_e - (V_e - V_0)e^{-t/RC}$. In this description, voltages are given in volts, capacitances in μf , and for resistances “K” means 1,000 ohms and “M” means 10^6 ohms.

The op amp A_2 is wired as shown to provide hysteresis [5]: its output snaps to positive voltage when the voltage on capacitor C_3 falls below ~ 2.015 and snaps to

ground when the voltage on C_3 rises above ~ 2.11 . Sensing begins when the former happens: switch S_2 connects the sensor part of the circuit to storage (capacitor C_1) and nearly instantaneously charges C_2 to approximately the same voltage $x(t_0)$ as is stored in C_1 . This causes the voltage in C_3 to rise past ~ 2.11 and flip S_2 to disconnect the sensor from C_1 . Then C_2 and C_3 , with combined capacitance 0.2, split their combined voltage $x(t_0) + 2.11$ and then discharge through resistor R_2 to return to sense C_1 again: by the RC-formula, the voltage on C_3 at time t is $(x(t_0) + 2.11/2)e^{-(t-t_0)/(10 \cdot 0.2)}$ and this reaches 2.015 when $t = t_0 + 2 \ln((x(t_0) + 2.11)/4.03)$.

Op amp A_1 controls the flow to and from C_1 : if the voltage on C_3 exceeds $s_d \approx 3.23$, the output of A_1 becomes positive and flips switch S_1 to discharge C_1 through the potentiometer R , otherwise C_1 charges through R_1 . Thus, if $x(t_0) > 2 \times 3.23 - 2.11 = 4.35$, capacitor C_1 discharges for the length of time $2 \ln(x(t_0) + 2.11/(2 \cdot 3.23))$ and then charges for the remaining time of the cycle, otherwise C_1 charges for the entire $2 \ln((x(t_0) + 2.11)/4.03)$. The formula for F follows directly from this and the RC-formula.

REFERENCES

1. *Chaos and Fractals, The Mathematics Behind the Computer Graphics*, R. L. Devaney and L. Keen, ed., AMS, Providence (1989).
2. P. Collet and J.-P. Eckmann, *Iterated maps of the interval as dynamical systems*, Birkhäuser, Boston (1980).
3. R. L. Devaney, *An Introduction to Chaotic Dynamical Systems*, Benjamin / Cummings, Reading (1986).
4. J. Guckenheimer and P. Holmes, *Nonlinear oscillations, dynamical systems, and bifurcation of vector fields*, Springer-Verlag, New York, 1983.
5. D. Lancaster and H. M. Berlin, *CMOS Cookbook*, SAMS, Carmel (1988).

Department of Mathematics and Computer Science
Mills College
Oakland, CA 94613
bassein@mills.edu

An Identity

$$(1-x)g(0) + xg(1) - g(x) = (1-x) \int_0^x tg''(t) dt + x \int_x^1 (1-t)g''(t) dt$$

Submitted by Michael Hirschhorn
Department of Pure Mathematics
The University of South Wales
Kensington, New South Wales
AUSTRALIA 2033

On the Geometry of Halley's Method

T. R. Scavo and J. B. Thoo

According to Traub [Tra64], Halley's iteration function (I.F.) "must share with the secant I.F. the distinction of being the most frequently rediscovered I.F. in the literature." Halley's method is a close relative of Newton's method, an iterative technique depicted as a sequence of tangent lines with zeros converging to a root of a function. The usual derivation of Halley's method, however, lacks any obvious geometric interpretation. We present a derivation of Halley's method having such an interpretation, and give a brief history of Halley's work and the method that bears his name.

1. HISTORICAL BACKGROUND. Edmond Halley (1656–1742), well-known astronomer and mathematician, was impressed by the work of "an ingenious professor of mathematics," Thomas Fautet de Lagny, who, in a book published in Paris in 1692,¹ presented some formulas for "extracting roots of pure powers, especially the cubic." Halley sought to understand the origin of these formulas, and in the process came to generalize them.

The result of de Lagny that impressed Halley² is that $\sqrt[3]{a^3 + b}$ lies between

$$a + \frac{ab}{3a^3 + b} \quad \text{and} \quad \frac{a}{2} + \sqrt{\frac{a^2}{4} + \frac{b}{3a}} \quad (1)$$

for $a^3 \gg b > 0$. Halley called these the *rational formula* and the *irrational formula*, respectively [Hal1694]. Each is a special case of more general iteration functions derived in Section 3.³

It is ironic that Halley preferred the irrational formula over the rational formula, for it is the latter that bears his name. Indeed, virtually all of Halley's calculations employed the irrational formula, of which he wrote [Hal1694]

And this formula is deservedly preferred before the rational one, which, on account of its large divisor, cannot be used without much trouble, in comparison of the irrational one, as manifold experience has informed me.

Apparently, extracting roots was relatively easy for Halley who claimed, for example, to have calculated eighteen significant digits of the cube root of 231 "in an hour's time" using the irrational formula.

Another reason Halley preferred the irrational formula was his belief that it generally gives better approximations than the rational formula. Speaking of the

¹See [Bat38] for a complete reference.

²de Lagny also gave a fifth-order formula that Halley found even *more* impressive.

³The formulas in (1) may be obtained by setting $f(x) := x^3 - (a^3 + b)$ in Equations (11) and (8), respectively, and evaluating at $x := a$.

methods in (1) he said

And between these two limits always lies the true root, being rather nearer to the irrational than to the rational formula

While true of each example given in [Hal1694], this is not true in general, however. A counterexample is provided in Section 4.

Halley also admired the work of the 16th century French mathematician François Viète, popularizer of what later became known as “Horner’s method,” an approximation technique pioneered by several Chinese mathematicians in the 13th century [Boy68]. Viète’s method, as it is sometimes called, is a linearly converging algorithm akin to bisection and may be applied to any polynomial with at least one real root [Ypm93]. It may also be used to produce starting values for Newton’s method and other higher-order iterative procedures, something that Halley himself might have done. More importantly, it appears that Viète’s method was an important precursor of Halley’s method, and of root-finding methods in general.

Although Halley was almost certainly aware of the fledgling calculus when he wrote his paper in 1694,⁴ he apparently did not realize that his method involved derivatives or fluxions, as he would have called them. In hindsight, this is not surprising considering it was Simpson, in 1740, who first realized the connection between derivatives and Newton’s method [Kol92]. What *is* surprising is that Brook Taylor recognized the derivatives in Halley’s method as early as 1712 [Fei85]:

[Taylor] noticed what Halley had failed to realize before him: that the coefficients in [Halley’s examples] are directly related to the successive derivatives of the original polynomial

Moreover, applying Halley’s techniques to Kepler’s problem—an outstanding problem in astronomy with which Halley was no doubt familiar—led Taylor to a remarkable discovery [Bai89, Ypm93]. In a letter to Machin in 1712, Taylor proclaimed [Fei85]

While I was thinking of these things, I fell into a general method of applying Dr. Halley’s Extraction of roots to all Problems . . . And it is comprehended in this Theorem

which turns out to be Taylor’s Theorem!

The reason that Kepler’s problem went unsolved for so long is that it involves a transcendental equation. In a superb summary of Halley’s work, Bateman [Bat38] suggests that Halley might have preceded Taylor in the discovery of Taylor’s formula had he only “applied his methods in a general way to transcendental equations.” While not important in and of itself—after all, Gregory knew of “Taylor’s theorem” around 1668—it is noteworthy that it was Halley’s method that prompted these developments, whereas Newton’s method languished in ignorance until the time of Simpson.

Despite Taylor’s achievements, he was unable to provide a general formula for Halley’s method. It remained for Schröder [Sch1870], more than one-and-a-half centuries later, to derive Halley’s iteration function as we now know it. But

⁴Newton published his *Principia* in 1687, but “only after intense coaxing” by Halley [Boy68]. In fact, the well-to-do Halley had the *Principia* published at his own expense.

Schröder made no reference to Halley. Indeed, Schröder was primarily interested in higher-order iteration functions and mentioned Halley’s formula almost in passing.

Kobald [Kob1891] derived Halley’s formula in a brief paper published in 1891, but unfortunately his derivation is unclear. Frame [Fra44], on the other hand, was the first to derive Halley’s iteration function via a second-degree Taylor polynomial (see Section 3 and also [Wal48, Ste51, Gan85, Bai89]). Some textbooks also employ this method (e.g., [Mcc67]), while others simply mention it.

Some authors have used determinants and Cramer’s rule to derive Halley’s formula and other higher-order iteration functions (see [Ham50, Ste51, Kis54]). Other derivations provided by Frame [Fra53] and Traub [Tra61b] used continued fractions and Padé approximants, respectively, while Snyder [Sny55] employed a technique he called the method of replacement. Finally, Salehov [Sal52] introduced the method of tangent hyperbolas discussed in Section 4.

2. PRELIMINARIES. Given a function $F: X \rightarrow X$ and a point $x_0 \in X$, one may iterate F to generate the sequence of points $x_0, x_1 = F(x_0), x_2 = F(x_1)$, and so forth. The sequence thus obtained,

$$x_{n+1} = F(x_n) \quad \text{for } n = 0, 1, 2, \dots, \quad (2)$$

is called the *orbit* of x_0 under iteration of F . A point α is called a *fixed point* of F if $F(\alpha) = \alpha$. When X is a subset of the real numbers, the graph of F intersects the line $y = x$ at each fixed point (see Figure 1). A fixed point α is said to be *attracting* if there exists a neighborhood U of α such that the orbit of every point $x_0 \in U$ converges to α under iteration of F . Finally, if $|F'(\alpha)| < 1$, then α is an attracting fixed point, an important result known over a century ago [Sch1870].

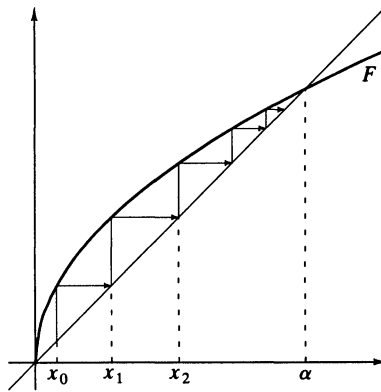


Figure 1. An orbit converging to a fixed point.

Now consider the problem of finding a root α of the equation

$$f(x) = 0. \quad (3)$$

We assume throughout that the root in question is *simple*, that is, $f'(\alpha) \neq 0$. One way to approximate α is to find another function F , called an *iteration function* (I.F.) for f , for which α is an attracting fixed point. Then, for a suitably chosen initial value x_0 , the iteration (2) converges to α . Note that the choice of I.F. is not unique (e.g., [Bur89, page 42]).

One well-known iterative root-finding method is *Newton's method*,

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \quad (4)$$

a special case of (2) with $F(x) = x - f(x)/f'(x)$. Evidently α is a fixed point of F if α is a simple root of (3); furthermore, this fixed point is attracting (see below). We call F the *Newton I.F.* for f , and denote it by N_f .

To derive (4), we approximate the given function f at $x = x_n$ by a linear function y of the form

$$y(x) = a(x - x_n) + b.$$

Then the requirement that both f and y , and their first derivatives, agree at $x = x_n$ leads to

$$y(x) = f'(x_n)(x - x_n) + f(x_n). \quad (5)$$

Finally, solving $y(x_{n+1}) = 0$ for x_{n+1} yields (4). Since (5) is the equation of the line tangent to f at $x = x_n$, it is clear that Newton's method applied to f may be interpreted as a sequence of tangent lines with zeros converging to a root of the function. (See Figure 2.)

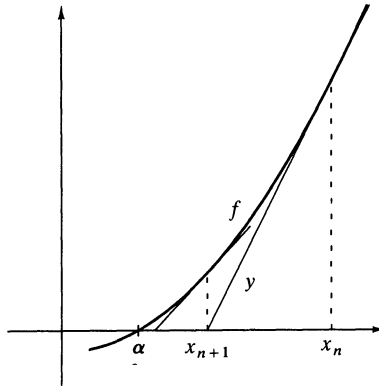


Figure 2. A geometric interpretation of Newton's method.

Newton's method is a quadratically converging root-finding algorithm. Loosely speaking, this means that the number of significant digits eventually doubles with each iteration. Such a method gives rise to a *second-order algorithm*. It can be shown that the first derivative of a second-order I.F. vanishes at the corresponding fixed point. In the case of Newton's I.F., the first derivative is

$$N'_f(x) = \frac{f(x)f''(x)}{f'(x)^2},$$

which clearly vanishes at a simple root α . Hence α is an attracting fixed point. And since $N'_f(\alpha) = f''(\alpha)/f'(\alpha)$ is nonzero in general, there exists a neighborhood for which (4) converges quadratically to α .

Whereas Newton's method is second order, we show in Section 3 that Halley's method,

$$x_{n+1} = x_n - \frac{2f(x_n)f'(x_n)}{2f'(x_n)^2 - f(x_n)f''(x_n)},$$

is a *third-order algorithm*. Such an algorithm converges cubically insofar as the number of significant digits eventually triples with each iteration. And not only does the first derivative of a third-order I.F. vanish at a fixed point, but so does the second derivative.

3. HALLEY'S METHOD. In Section 2 we derived Newton's method using a linear function y , the first-degree Taylor polynomial of f at x_n . Let's see what happens if we instead use a second-degree Taylor polynomial,

$$y(x) = f(x_n) + f'(x_n)(x - x_n) + \frac{f''(x_n)}{2}(x - x_n)^2,$$

where x_n is again an approximate root of $f(x) = 0$. As with Newton's method, the goal is to determine a point x_{n+1} where the graph of y intersects the x -axis, that is, to solve the equation

$$0 = f(x_n) + f'(x_n)(x_{n+1} - x_n) + \frac{f''(x_n)}{2}(x_{n+1} - x_n)^2 \quad (6)$$

for x_{n+1} .⁵ Following Frame [Fra44] and others, we factor $x_{n+1} - x_n$ from the last two terms of (6) to obtain

$$0 = f(x_n) + (x_{n+1} - x_n) \left(f'(x_n) + \frac{f''(x_n)}{2}(x_{n+1} - x_n) \right),$$

from which it follows that

$$x_{n+1} - x_n = - \frac{f(x_n)}{f'(x_n) + \frac{f''(x_n)}{2}(x_{n+1} - x_n)}. \quad (9)$$

Approximating the difference $x_{n+1} - x_n$ remaining on the right-hand side of (9) by Newton's correction $-f(x_n)/f'(x_n)$ given in (4), we obtain

$$x_{n+1} = x_n - \frac{2f(x_n)f'(x_n)}{2f'(x_n)^2 - f(x_n)f''(x_n)}, \quad (10)$$

widely known as *Halley's method* [Bat38, Ste51, Fra53, Kis54, Sny55, Tra61b, Tra64, Dav75, Bro77, Han77, Pop80, Ale81, Gan85].

Unfortunately, the preceding derivation lacks any clear geometric interpretation analogous to the tangent lines of Newton's method. What we seek is an *osculating curve* to f at x_n (that is, a curve agreeing with f at x_n up through second derivative) that interpolates the points $(x_n, f(x_n))$ and $(x_{n+1}, 0)$ where x_{n+1} is

⁵One might be tempted to apply the quadratic formula to (6), obtaining

$$x_{n+1} - x_n = \frac{-f'(x_n) \pm \sqrt{f'(x_n)^2 - 2f(x_n)f''(x_n)}}{f''(x_n)}. \quad (7)$$

A judicious choice of sign in (7) (see [Tra64, Gor90]) leads to the I.F.

$$C_f(x) = x - \frac{1 - \sqrt{1 - 2f(x)f''(x)/f'(x)^2}}{f''(x)/f'(x)}, \quad (8)$$

a general form of Halley's irrational formula [Hal1694, Bat38, Gan85]. But neither (7) nor (8) is what is known as Halley's method. We remark that rationalizing the numerator of (7) yields a special case of Laguerre's method [Ost73, Han77] sometimes attributed to Cauchy [Tra64, Pop80].

given in (10). If the curve crosses the x -axis but once, so much the better. This brings us to the so-called “method of tangent hyperbolas,” but first we make a few remarks concerning Halley’s method.

Denote the *Halley I.F.* for f by

$$H_f(x) = x - \frac{2f(x)f'(x)}{2f'(x)^2 - f(x)f''(x)}. \quad (11)$$

If α is a simple zero of f , then we see immediately that $H_f(\alpha) = \alpha$. Further, a straightforward calculation shows that $H'_f(\alpha) = H''_f(\alpha) = 0$ while $H'''_f(\alpha) \neq 0$. Thus, Halley’s I.F. is third order for simple roots. In fact, a direct computation shows that

$$H'''_f(\alpha) = - \left(\frac{f'''(\alpha)}{f'(\alpha)} - \frac{3}{2} \left(\frac{f''(\alpha)}{f'(\alpha)} \right)^2 \right) = -Sf(\alpha),$$

where $Sf(x)$ denotes the Schwarzian derivative of f at x , a most curious result.⁶

Bateman [Bat38] was the first to point out that Halley’s method may be obtained by applying Newton’s method to $f/\sqrt{f'}$, that is,

$$N_{f/\sqrt{f'}} = x - \frac{f/\sqrt{f'}}{(f/\sqrt{f'})'} = H_f.$$

And despite the uncanny similarity, Halley’s method is not to be confused with a second-order method for multiple roots discovered by Schröder,

$$x_{n+1} = x_n - \frac{f(x_n)f'(x_n)}{f'(x_n)^2 - f(x_n)f''(x_n)},$$

obtained by applying Newton’s method to f/f' [Sch1870, Bod49, Tra64, Bur89].

The following special case of Halley’s method is also worth investigating. Let $g(x) = x^d - r$. Then, by (11), the Halley I.F. for g is

$$H_g(x) = \frac{(d-1)x^d + (d+1)r}{(d+1)x^d + (d-1)r}x, \quad (12)$$

a result often ascribed to Bailey [Bai41, Fra45, Tra61a], but actually due to Lambert in 1770 [Kis54, Tra61b]. Traub [Tra64] remarks that some early authors called (12) “Hutton’s method” without reference. Indeed, a footnote in the English translation of Halley’s paper [Hal1694, page 644] specifically attributes (12) to Hutton in 1786,⁷ but this clearly postdates Lambert’s work. (See [Bat38, Wal48] and especially [Bai89] for more information on Lambert’s method.)

We close this section with a graphical example comparing the methods of Newton and Halley. Let $f(x) = x^2 - 2$. Then

$$N_f(x) = \frac{x^2 + 2}{2x} \quad \text{and} \quad H_f(x) = \frac{x^3 + 6x}{3x^2 + 2}.$$

Since $f(\sqrt{2}) = 0$, it follows that $N_f(\sqrt{2}) = H_f(\sqrt{2}) = \sqrt{2}$ (see Figure 3a). Moreover, this fixed point is attracting since $N'_f(\sqrt{2}) = H'_f(\sqrt{2}) = 0$. And because the second derivative of H_f also vanishes, whereas the second derivative of N_f does not, the graph of H_f is flatter than that of N_f near the fixed point (see Figure 3b). This accounts for the difference in speed at which the two algorithms converge

⁶The Schwarzian derivative is an important tool in the study of discrete dynamical systems [Dev92].

⁷Hutton was one of the translators.

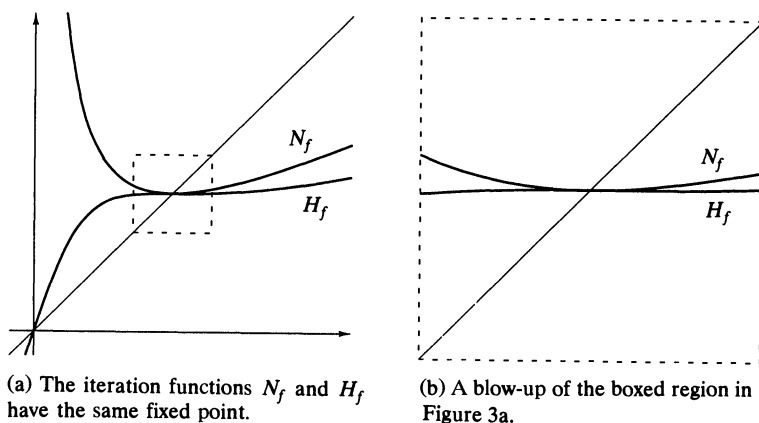


Figure 3. Typical Newton and Halley iteration functions.

(see [Wal48, Bod49] for details). In general, the higher the order, the flatter the graph and, hence, the faster the convergence.

4. THE METHOD OF TANGENT HYPERBOLAS. Salehov [Sal52] was apparently the first to suggest that Halley's I.F. could be derived using an osculating rational function of the form

$$y(x) = \frac{x + c}{ax + b}. \quad (13)$$

(Recall from page 421 that an osculating curve to f at x_n is one that satisfies the equations

$$y^{(k)}(x_n) = f^{(k)}(x_n) \quad (14)$$

for $k = 0, 1, 2$.) For convenience, we use an equivalent form of (13),

$$y(x) = \frac{(x - x_n) + c}{a(x - x_n) + b}. \quad (15)$$

Equations (14) and (15) taken together lead to the system of equations

$$\begin{cases} \frac{c}{b} = f(x_n) \\ \frac{b - ac}{b^2} = f'(x_n) \\ \frac{2a(ac - b)}{b^3} = f''(x_n) \end{cases}$$

having solution

$$\begin{cases} a = \frac{-f''(x_n)}{2f'(x_n)^2 - f(x_n)f''(x_n)} \\ b = \frac{2f'(x_n)}{2f'(x_n)^2 - f(x_n)f''(x_n)} \\ c = \frac{2f(x_n)f'(x_n)}{2f'(x_n)^2 - f(x_n)f''(x_n)} \end{cases}. \quad (16)$$

It follows from (15) that if $y(x_{n+1}) = 0$, then $x_{n+1} = x_n - c$ where c is given in (16), that is,

$$x_{n+1} = x_n - \frac{2f(x_n)f'(x_n)}{2f'(x_n)^2 - f(x_n)f''(x_n)}.$$

But this is precisely Halley’s method given in (10). In other words, Halley’s formula can be derived using an osculating hyperbola. Indeed, Halley’s method is sometimes called the *method of tangent hyperbolas* [Sal52, Saf63, Tra64].

As an example, consider the function $f(x) = e^x - 2$ which has a unique zero at $\alpha = \log 2$. The Halley I.F. for f is

$$H_f(x) = x - \frac{2(e^x - 2)e^x}{2e^{2x} - (e^x - 2)e^x} = \frac{(x - 2)e^x + 2(x + 2)}{e^x + 2}.$$

Observe that the graph of H_f is asymptotic to the diagonal lines $y = x \mp 2$ as $x \rightarrow \pm \infty$, respectively (see Figure 4a). Indeed, a direct calculation shows that

$$0 \leq H'_f(x) = \left(\frac{e^x - 2}{e^x + 2} \right)^2 < 1$$

for all x , making the fixed point globally attracting. Consequently, we may choose any initial value we please. For instance, using the starting value $x_0 = 10$, system (16) yields approximately

$$\begin{cases} a = -4.539580784 \times 10^{-5} \\ b = 9.079161568 \times 10^{-5} \\ c = 1.999636834 \end{cases} \tag{17}$$

from which we obtain $x_1 \approx x_0 - c = 8.000363166$. Similarly, substituting $x_1 = 8.000363166$ in (16) gives approximately

$$\begin{cases} a = -3.351160651 \times 10^{-4} \\ b = 6.702321302 \times 10^{-4} \\ c = 1.997319071 \end{cases} \tag{18}$$

Thus, $x_2 \approx x_1 - c = 6.003044095$.

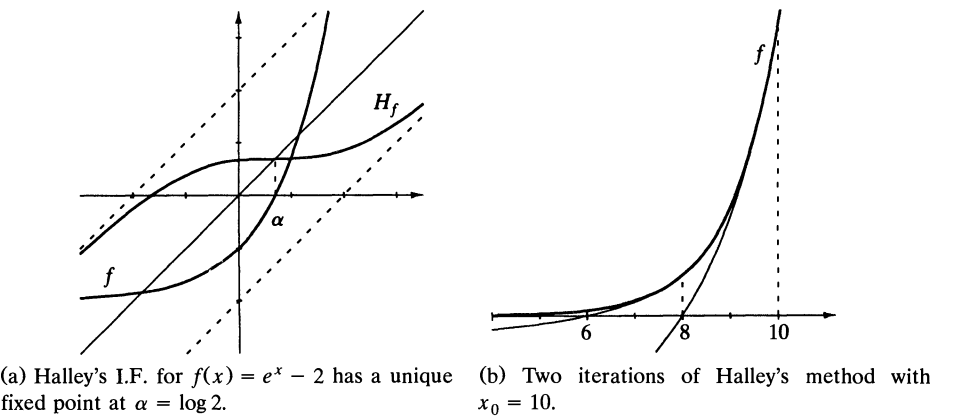


Figure 4. A visualization of Halley’s method.

The osculating hyperbolas (15) corresponding to (17) and (18) are plotted alongside the graph of f in Figure 4b. Notice that upon successive applications of Halley's method, the zeros of the tangent hyperbolas tend to the zero of f . This is the soughtafter geometric interpretation of Halley's method. Incidentally, continuing this process numerically, we find that x_7 agrees with α to ten decimal places, and thereafter the number of significant digits roughly triples with each iteration.

The function $f(x) = e^x - 2$ also provides a counterexample to Halley's claim that the irrational formula is generally better than the rational formula. Observe that the first two points on the orbit of $x_0 = 1.3$ under iteration of C_f given in (8) are

$$1.3 \mapsto 0.60021187 \dots \mapsto 0.69327247 \dots,$$

whereas Halley's method gives

$$1.3 \mapsto 0.71110978 \dots \mapsto 0.69314766 \dots.$$

Since $\log 2 = 0.69314718 \dots$, we see that Halley's method gives better approximations in this case. Thus, contrary to Halley's claim, the irrational formula does not always give better approximations than the rational formula.

ACKNOWLEDGMENTS. We thank M. Iklé for comments on an early draft of this paper, and G. D. Chakerian and B. N. Parlett for critical readings of later versions. Chakerian also pointed out a faulty geometric interpretation in [Fra44]. A. Zorich graciously provided a translation of [Sal52], while G. W. Stewart made available a detailed translation of Schröder's seminal paper [Sch1870]. Last, but not least, we thank the referees for their helpful suggestions, and for alerting us to an error in the original manuscript.

Postscript. After this paper was accepted for publication, the authors learned from W. Gander that he gave a geometric interpretation of Halley's method a decade earlier which was deleted from the published version of his manuscript [Gan85].

REFERENCES

-
- [Hal1694] Halley, Edmund (1694). A new, exact, and easy method of finding the roots of any equations generally, and that without any previous reduction (Latin). *Philos. Trans. Roy. Soc. London* 18, 136–148. [English translation: *Philos. Trans. Roy. Soc. London* (abridged) 3, 640–649 (1809).]
 - [Sch1870] Schröder, E. (1870). On infinitely many algorithms for solving equations (German). *Math. Ann.* 2, 317–365. [English translation by G. W. Stewart, TR-92-121, Institute for Advanced Computer Studies, University of Maryland, College Park, MD, 20742].
 - [Kob1891] Kobald, E. (1891). Notice concerned with the calculation of roots of numerical equations (German). *Monatsh. Math. und Physik* 2, 331–332.
 - [Bat38] Bateman, H. (1938). Halley's methods for solving equations. *Amer. Math. Monthly* 45, 11–17.
 - [Bai41] Bailey, V. A. (1941). Prodigious calculation. *Austral. J. Sci.* 3(4), 78–80.
 - [Fra44] Frame, J. S. (1944). A variation of Newton's method. *Amer. Math. Monthly* 51, 36–38.
 - [Fra45] Frame, J. S. (1945). Remarks on a variation of Newton's method. *Amer. Math. Monthly* 52, 212–214.
 - [Wal48] Wall, H. S. (1948). A modification of Newton's method. *Amer. Math. Monthly* 55, 90–94.
 - [Bod49] Bodewig, E. (1949). On types of convergence and on the behavior of approximations in the neighborhood of a multiple root of an equation. *Quart. Appl. Math.* 7, 325–333.
 - [Ham50] Hamilton, H. J. (1950). A type of variation on Newton's method. *Amer. Math. Monthly* 57, 517–522.
 - [Ste51] Stewart, J. K. (1951). Another variation of Newton's method. *Amer. Math. Monthly* 58, 331–334.
 - [Sal52] Salehovich, G. S. (1952). On the convergence of the process of tangent hyperbolas (Russian). *Dokl. Akad. Nauk SSSR* 82, 525–528.
 - [Fra53] Frame, J. S. (1953). The solution of equations by continued fractions. *Amer. Math. Monthly* 60, 293–305.

- [Kis54] Kiss, I. (1954). A generalization of Newton's approximation procedure (German). *Z. Angew. Math. Mech.* 34, 68–69.
- [Sny55] Snyder, R. W. (1955). One more correction formula. *Amer. Math. Monthly* 62, 722–725.
- [Tra61a] Traub, J. F. (1961). Comparison of iterative methods for the calculation of n th roots. *Comm. ACM* 4(3), 143–145.
- [Tra61b] Traub, J. F. (1961). On a class of iteration formulas and some historical notes. *Comm. ACM* 4(6), 276–278.
- [Saf63] Šafiev, R. A. (1963). The method of tangent hyperbolas. *Soviet Math. Dokl.* 4, 482–485. [Russian original: *Dokl. Akad. Nauk SSSR* 149, 788–791 (1963).]
- [Tra64] Traub, J. F. (1964). *Iterative Methods for the Solution of Equations*. New Jersey: Prentice-Hall.
- [McC67] McCalla, T. R. (1967). *Introduction to Numerical Methods and FORTRAN Programming*. New York: Wiley.
- [Boy68] Boyer, C. B. (1968). *A History of Mathematics*. New York: Wiley.
- [Ost73] Ostrowski, A. M. (1973). *Solution of Equations in Euclidean and Banach Spaces*. New York: Academic Press. [Third edition of *Solution of Equations and Systems of Equations* by the same author.]
- [Dav75] Davies, M. and B. Dawson (1975). On the global convergence of Halley's iteration formula. *Numer. Math.* 24, 133–135.
- [Bro77] Brown, G. H., Jr. (1977). On Halley's variation of Newton's method. *Amer. Math. Monthly* 84, 726–728.
- [Han77] Hansen, E. and M. Patrick (1977). A family of root finding methods. *Numer. Math.* 27, 257–269.
- [Pop80] Popovski, D. B. (1980). A family of one-point iteration formulae for finding roots. *Internat. J. Comput. Math.* 8, 85–88.
- [Ale81] Alefeld, G. (1981). On the convergence of Halley's method. *Amer. Math. Monthly* 88, 530–536.
- [Gan85] Gander, W. (1985). On Halley's iteration method. *Amer. Math. Monthly* 92, 131–134.
- [Fei85] Feigenbaum, L. (1985). Taylor and the method of increments. *Arch. Hist. Exact Sci.* 34, 1–140.
- [Bur89] Burden, R. L. and J. D. Faires (1989). *Numerical Analysis* (fourth edition). Boston: PWS-Kent.
- [Bai89] Bailey, D. F. (1989). A historical survey of solution by functional iteration. *Math. Mag.* 62(3), 155–166.
- [Gor90] Gordon, S. P. and E. R. von Eschen (1990). A parabolic extension of Newton's method. *Internat. J. Math. Ed. Sci. Tech.* 21(4), 519–525.
- [Dev92] Devaney, R. L. (1992). *A First Course in Chaotic Dynamical Systems: Theory and Experiment*. Reading, MA: Addison-Wesley.
- [Kol92] Kollerstrom, N. (1992). Thomas Simpson and 'Newton's method of approximation': an enduring myth. *Brit. J. Hist. Sci.* 25, 347–354.
- [Ypm93] Ypma, T. J. (1993). Historical development of the Newton-Raphson-Simpson method. Preprint.

616 Westcott Street
Syracuse, NY 13210
trscavo@mailbox.syr.edu

Department of Mathematics
University of California
Davis, CA 95616-8633
jb2@math.ucdavis.edu

Answer to Picture Puzzle (p. 408)

This is a portrait of Georg Cantor in his late twenties, around the early 1870s when he was starting his work on set theory and transfinite arithmetic. Few photographs of Cantor from this period of his life seem to exist, and none have previously been published.

The lady in the photograph is his sister Sophie, three years his junior. She was important through his life. In particular, when he suffered mental illness from 1899, he would sometimes stay with her and her family after release from mental hospitals or sanatoria.

We thank Ivor Grattan-Guinness for providing the photograph.

The Binary Expansion of $\frac{1}{p}$

A. R. Meijer

Given a recurring sequence s_0, s_1, s_2, \dots consisting of 0s and 1s, it is always possible to find a recurrence relation of the form

$$s_{n+k} = a_{k-1}s_{n+k-1} + a_{k-2}s_{n+k-2} + \dots + a_1s_{n+1} + a_0s_n \quad (1)$$

which if s_0, s_1, \dots, s_{k-1} are used as initial values, will generate all the subsequent terms in the sequence. (The a_i in (1) are elements of the field $\text{GF}(2) = \mathbf{Z}/2\mathbf{Z}$, and the operations are assumed to be in that field.) For example, if the sequence has period t , it can always be generated, in a rather trivial way, by the recurrence relation

$$s_{n+t} = s_n \quad (2)$$

using the terms s_0, s_1, \dots, s_{t-1} as initial values.

It is customary to associate with the relation (1) the polynomial

$$h(x) = x^k + a_{k-1}x^{k-1} + \dots + a_1x + a_0 \quad (3)$$

which is said to *generate* the sequence. It might be better to view (3) as an annihilator of the sequence, in the sense that, given any segment $s_i, s_{i+1}, \dots, s_{i+k}$, where $k = \text{degree of } h(x)$ and i is any natural number, one has

$$s_{i+k} + a_{k-1}s_{i+k-1} + \dots + a_1s_{i+1} + a_0s_i = 0.$$

The family of all polynomials which annihilate the sequence in this sense is easily shown to be an ideal in the principal ideal domain $\text{GF}(2)[x]$, and consequently there exists a polynomial $g(x)$ (of minimal degree in this ideal) such that every annihilator of the sequence is a multiple of $g(x)$. In particular (2) shows that if the sequence has period t , then $x^t + 1$ is a multiple of $g(x)$. The degree of $g(x)$ is called the *linear complexity* of the given sequence. An efficient algorithm, due to Berlekamp and Massey exists for finding $g(x)$. (See, for example, [1, p. 176].)

In this note we find $g(x)$ in the particular case where the given sequence is the binary expansion of $\frac{1}{p}$, where p is a prime such that 2 is primitive modulo p : thus no power of 2 less than 2^{p-1} is congruent to 1 modulo p , or, equivalently, the binary expansion is recurring with period $p-1$. In many applications one would wish p to be large (of order 10^{100} to 10^{150}) in which case use of the Berlekamp-Massey algorithm is infeasible.

We shall restrict ourselves to primes p of the form $p = 2q + 1$, where q is itself a prime, $q \equiv 1$ modulo 4. Standard results in number theory (see [6], for example) then guarantee that 2 is primitive modulo p . It should be pointed out, however, that the final result of this note holds for any p with 2 primitive, and only minor modifications to the argument are necessary to prove this. Primes of the form that we consider ("safe primes" under a rather loose definition of "safety" for use in the RSA cryptosystem) have the advantage of being relatively easy to find [4], even though it is unknown whether an infinite number of them exist.

By way of a very small example, let us consider the case $p = 11$. Then

$$\begin{aligned}
 \frac{1}{11} &= 0.0001011101\dots \\
 \frac{2}{11} &= 0.0010111010\dots \\
 \frac{4}{11} &= 0.0101110100\dots \\
 \frac{8}{11} &= 0.1011101000\dots \\
 \frac{5}{11} &= 0.0111010001\dots \\
 \frac{10}{11} &= 0.1110100010\dots \\
 \frac{9}{11} &= 0.1101000101\dots \\
 \frac{7}{11} &= 0.1010001011\dots \\
 \frac{3}{11} &= 0.0100010111\dots \\
 \frac{6}{11} &= 0.1000101110\dots
 \end{aligned} \tag{4}$$

Our interest in the expansion of $\frac{1}{p}$ stems from the fact that the sequence

$$\frac{1}{p} = 0.s_0s_1s_2s_3\dots a_4\dots \tag{5}$$

can be shown to display “good” pseudo-randomness properties—“good” in the sense that it looks pretty much like a really random sequence of 0s and 1s. (Postulates for “goodness” were laid down by S. W. Golomb in [3].) The sequence has in fact been studied for this reason by Blum, Blum and Shub [2]. Moreover, given p , generating the sequence (5) is extremely fast, which would seem to recommend its use in, for example, bit stream ciphers. (In a bit stream cipher a message is made unintelligible to anyone not in possession of a secret key, by adding a pseudo-random bit, modulo 2, to every bit of the message.) If one observes, as in the example above, that for any integer $x \in [1, p - 1]$ the expansion of (x/p) is a cyclic shift of the expansion of $\frac{1}{p}$, this becomes even more

tempting, since x can then be used as the key. Regrettably, as shown in [2], such a cipher would be very insecure against attack, since only $\lceil \log_2 p \rceil$ consecutive bits determine x completely: a fact which is easily seen if these $\lceil \log_2 p \rceil$ bits are the first ones; if the sequence is $b_1b_2b_3\dots$, $b_i \in \{0, 1\}$, then x is the integer nearest

$$p \times \left(\frac{b_1}{2} + \frac{b_2}{4} + \frac{b_3}{8} + \dots \right).$$

The general case, in which the known bits appear somewhere in the middle, merely involves a cyclic shift of this special case.

For a general discussion of bit stream ciphers and the linear feedback shift registers (LFSRs) generating pseudo-random bit sequences, we refer the reader to

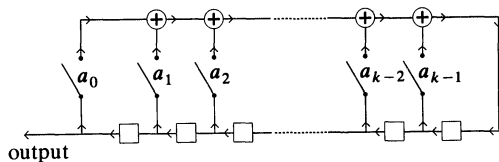


Figure 1

[7] or [5]. An LFSR may be considered to be an implementation in hardware of the recurrence relation (1), as in figure 1, in which the square boxes represent single bit memory registers (“flip-flops”), the switches represent the coefficients a_i (i.e. if $a_i = 0$ the i th switch is open, if $a_i = 1$, it is closed) and \oplus denotes addition modulo 2. At each clock pulse, the contents of the registers are shifted in the indicated direction. The reader may verify that if the registers in figure 2 initially contain 000101, then the output sequence will be the binary expansion of $\frac{1}{11}$.

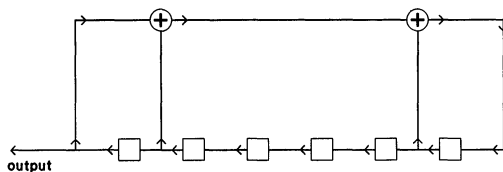


Figure 2

The linear complexity of the sequence, that is the minimal degree among all polynomials which generate the sequence, now translates into the minimal number of memory registers required to produce the sequence. Figure 2 shows that the sequence (4) has a linear complexity of at most 6, or equivalently, that (4) may be generated by the recurrence relation

$$s_{n+6} = s_{n+5} + s_{n+1} + s_n$$

with which we associate, as in (3), the polynomial

$$x^6 + x^5 + x + 1. \quad (6)$$

Linear complexity measures to some extent the pseudo-randomness of the sequence, but only in one direction: a “good” pseudo-random sequence must have a high linear complexity. (To see that the converse is false, consider the sequence consisting of 10^{80} zeros, followed by a one, and then recurring. This has linear complexity = period = $10^{80} + 1$, but does not appear very random!)

We shall show that, if p is of the form $p = 2q + 1$, with $q \equiv 1$ modulo 4 also prime, then the linear complexity of the binary expansion of $\frac{1}{p}$ is $\frac{1}{2}(p + 1)$.

To this end, note in the first place that, by Euler’s theorem, $2^{(p-1)/2} \equiv p - 1$ modulo p . Thus $\frac{1}{p}$ and the fractional part of $2^{(p-1)/2}/p$ add up to $1 = 0.11111\dots$ It follows immediately that the first $(p - 1)/2$ digits of $\frac{1}{p}$ and the second $(p - 1)/2$ digits are each other’s complements. This implies of course that the recurring part of $\frac{1}{p}$ contains equal numbers of 0s and 1s. It is also clear that the expansions of $\frac{2}{p}$ and of the fractional part of $2^{(p+1)/2}/p$ are similarly complementary (as shown in the example above by $\frac{1}{11}$ and $\frac{10}{11}$ and by $\frac{2}{11}$ and $\frac{9}{11}$). Thus, if we denote by c_i the expansion of the fractional part of $\frac{2^i}{p}$, then we have, adding componentwise modulo 2, that

$$c_{(p+1)/2} + c_{(p-1)/2} + c_1 + c_0 = 0$$

or, in the form of equation (1), that

$$s_{n+(p+1)/2} = s_n + s_{n+1} + s_{n+(p-1)/2}$$

or, in terms of a generating polynomial, that

$$h(x) = 1 + x + x^{(p-1)/2} + x^{(p+1)/2}$$

generates the sequence. The linear complexity of the $\frac{1}{p}$ sequence is therefore at most $(p+1)/2$.

To prove that it cannot be less than this, suppose that a relation of the form

$$a_0 \mathbf{c}_n + a_1 \mathbf{c}_{n+1} + \cdots + a_{k-1} \mathbf{c}_{n+k-1} + \mathbf{c}_{n+k} = \mathbf{0} \quad (7)$$

holds, with k minimal. Since the first $p-1$ components of \mathbf{c}_i contain equal numbers of 0s and 1s, it is clear that the number of nonzero terms in (7) must be even. Moreover, for $0 \leq k < p-1$, all the \mathbf{c}_i are distinct, so the number of nonzero terms in (7) must be greater than 2.

Next recall, that $g(x)$ must be a divisor of $x^{p-1} + 1$

$$\begin{aligned} &= (x^{(p-1)/2} + 1)^2 \\ &= (x+1)^2 [x^{(p-3)/2} + \cdots + x + 1]^2. \end{aligned}$$

Now the term inside the square parentheses has an odd number of terms (since $p \equiv 3$ modulo 8) and therefore it cannot be the generator polynomial nor, clearly, can any of its divisors. Obviously $g(x) \neq x+1$.

The product of $x+1$ and the term inside the square parentheses has only two terms, and is therefore also ineligible. The polynomial of least degree which might do the trick is therefore

$$\begin{aligned} &(x+1)^2 [x^{(p-3)/2} + \cdots + x + 1] \\ &= x^{(p+1)/2} + x^{(p-1)/2} + x + 1 \end{aligned}$$

and it follows that the linear complexity of the sequence (4) is therefore at least $(p+1)/2$, which completes the proof. In fact, putting these two parts of the proof together, we see that

$$g(x) = x^{(p+1)/2} + x^{(p-1)/2} + x + 1$$

is of least degree generating the sequence, of which (6), or the associated circuit of figure 2, is an example.

It is interesting that this method enables one to generate pseudo-random bit sequences of, apparently, arbitrarily high linear complexity; certainly of higher complexity than could ever be physically implemented in an LFSR.

REFERENCES

1. Blahut, R. E.: *Theory and Practice of Error Control Codes*, Addison-Wesley, Reading, MA, 1983.
2. Blum, L., Blum, M., and Shub, M.: A single unpredictable pseudo-random generator; *SIAM J. Comput.* 15 (1986).
3. Golomb, S. W.: *Shift Register Sequences*, Holden-Day, San Francisco, 1967.
4. Gordon, J. A.: Strong primes are easy to find; in Beth. T. et al. (eds.): *Advances in Cryptology: Proceedings Eurocrypt '84*, Lecture Notes in Computer Science 209; Springer Verlag, Berlin, 1985.
5. Tilborg, H. C. A. van: *An Introduction to Cryptology*; Kluwer, Norwell, MA, 1988.
6. Vinogradov, I. M.: *Elements of Number Theory*; Dover, New York, 1954.
7. Welsh, D.: *Codes and Cryptography*, Oxford University Press, New York, 1989.

Department of Mathematics & Applied Mathematics
University of Natal
King George V Avenue
Durban 4001, South Africa

Pick's Formula via the Weierstrass \wp -Function

Ricardo Diaz and Sinai Robins

1. INTRODUCTION. If the vertices of a polygon lie on the lattice of points in the plane whose coordinates are integers, then Pick's formula provides a method for computing the area of the polygon simply by counting lattice points. Let B denote the number of lattice points that lie on the edges of a simply closed polygon (including vertices), and let I denote the number of lattice points that lie within the interior of the polygon. The area A of the polygon, according to Pick, is

$$A = I + \frac{1}{2}B - 1.$$

This simple relation is comprehensible to a fifth grader [17], yet the theorem continues to intrigue modern researchers because of its deeper connections with combinatorics and algebraic geometry. A multitude of alternative proofs and generalizations of Pick's formula have appeared since the original proof in 1899 ([7], [8], [9], [11], [15], [25]). The first generalization to lattices in three dimension was made by J. E. Reeve in 1957 [19], who cleverly introduced an auxiliary lattice. His work was subsequently extended by others to higher dimensions as part of an analysis of the so-called 'Ehrhart polynomials' of polyhedra. ([4], [5], [12], [14], [18], [20], [22], [24]). Recently, connections between the algebraic-geometrical properties of toric varieties and the number of lattice points in polyhedra have been investigated ([2], [18]). There are also generalizations of Pick's formula to other types of archimedean lattices ([10], [21]); and to self-intersecting polygons ([10], [23], [26]). Although our complex-analytic proof of Pick's formula is not the shortest available (by any means!), it does expose a host of connections between lattice geometry, the Weierstrass \wp -function, classical magnetostatics, and Kodaira-Hodge-DeRham theory. Our approach is motivated by the idea that Pick's theorem is a discrete version of Green's theorem in the plane.

We first summarize for the reader some classical results in complex analysis pertaining to complex-valued functions that are doubly-periodic; that is, invariant with respect to translation by integral multiples of two linearly-independent vectors in the complex plane. Those readers already conversant with these results will realize that traditionally in complex analysis these vectors are not required to be perpendicular or of unit length; but for the purposes of this paper we found it convenient to adopt the simplifying convention that the two vectors correspond to 1 and i . The lattice they generate is called the Gaussian lattice. Of course a version of Pick's theorem holds for any lattice generated by two linearly independent vectors in the complex plane, since an affine transformation leaves the linear relationship of Pick's formula invariant except for a multiplicative change of scale for area.

Definition 1. A function $f(z)$ is said to be *doubly-periodic* with periods i and 1 if $f(z) = f(z + i) = f(z + 1)$. This implies that $f(z + m + in) = f(z)$ for all integers m, n .

If $f(z)$ is doubly-periodic, then the integral of $f(z) dz$ along opposite sides of a parallelogram cancel if the vertices of the parallelogram lie on the square lattice and if the sides carry opposing orientations. Note that despite the use of complex notation, we are not requiring here that $f(z)$ be an *analytic* function.

Definition 2. The Weierstrass \wp -function for the Gaussian lattice is defined by

$$\wp(z) = z^{-2} + \sum_a' \left[(z - a)^{-2} - a^{-2} \right]$$

in which the sum extends over all Gaussian lattice points except $a = (0, 0)$.

The sum converges uniformly and absolutely on compact subsets of the lattice-punctured plane. Because the sum is invariant under rearrangement of its terms, one can check easily that $\wp(z)$ is a doubly-periodic function which is analytic except at the lattice points. Because the lattice is invariant under multiplication by i , one can also verify from the preceding formula that $\wp(iz) = -\wp(z)$. The residue of $\wp(z)$ at each pole is zero, because all poles are double poles.

Definition 3. An antiderivative for the function $-\wp(z)$ is given by the Weierstrass zeta function

$$\zeta(z) = z^{-1} + \sum_a' \left[(z - a)^{-1} + a^{-1} + za^{-2} \right].$$

This series also converges uniformly and absolutely on compact subsets of the lattice-punctured plane. One can verify that $d\zeta(z)/dz = -\wp(z)$ by termwise-differentiation of the series defining $\zeta(z)$. Note that $\zeta(z)$ has its poles located precisely at all lattice points, and that the residue of $\zeta(z)$ at each pole is 1 . It is easy to verify that $\zeta(z)$ is an odd function. This follows from the fact that the lattice is invariant under multiplication by -1 . Unfortunately, $\zeta(z)$ is *not* doubly-periodic (but rather is ‘pseudo-periodic’ in the terminology of [13]).

§2. THE INGREDIENTS. The following Lemma shows that the Weierstrass ζ -function is only a conjugate-analytic, linear term away from being doubly periodic. We could show this by using the ‘Legendre-relations’ for ζ [13], but instead include a self-contained proof for completeness.

Lemma 1. *There exists a constant α such that the function $\phi(z) = \zeta(z) - \alpha \bar{z}$ is doubly-periodic with periods i and 1 .*

Proof: The obstruction to double-periodicity of $\zeta(z)$ is the nonvanishing of the expression

$$\zeta(z + m + in) - \zeta(z) = - \int_{w=0}^{w=m+in} \wp(z + w) dw.$$

The preceding definite integral is independent of path because the residues of $\wp(z)$ are zero. Because $\wp(z)$ is doubly-periodic, this integral can be expressed as the sum of m duplicates of an integration taken along a horizontal linear path of length one and n duplicates of an integration taken along a vertical linear path of length one. Thus the integral equals $m\alpha + n\beta$ where $\alpha(z) = - \int_0^1 \wp(z + w) dw$

and $\beta(z) = -\int_0^1 \wp(z + iw)idw$. Since the integrand in the definition of $\alpha(z)$ is periodic on each horizontal line, the quantity $\alpha(z)$ is invariant under horizontal translations of z ; hence $\alpha(z) = \alpha(y)$. Similarly $\beta(z) = \beta(x)$. However, the expression $\zeta(z + m + in) - \zeta(z) = m\alpha(y) + in\beta(x)$ must be analytic in the variable z on the lattice-punctured plane. Set $m = 0$ to deduce that $\beta(x)$ must be analytic on the lattice-punctured plane. But from the Cauchy-Riemann equations it follows that an analytic function that depends on only one of the real coordinates must be constant. Similarly α must be constant. Returning then to the definitions of α and β one sees upon integration of the identity $\wp(iz) = -\wp(z)$ that there is the relation $\beta = -i\alpha$. Thus $\zeta(z + m + in) - \zeta(z) = (m - in)\alpha = [(z + m + in) - \bar{z}]\alpha$. This implies that $\zeta(z) - \alpha\bar{z}$ is doubly-periodic. Q.E.D.

The doubly-periodic function $\phi(z)$ is not analytic because of the presence of the conjugate-analytic, linear term $\alpha\bar{z}$. Nevertheless, $\phi(z)$ has some very nice integral properties. Let C denote a canonically oriented curvilinear polygonal path in the complex-plane; that is, a piecewise continuously-differentiable, simple closed curve that winds once counterclockwise around a bounded, simply-connected domain D .

Lemma 2. *If C passes through no lattice points, then the number I of lattice points inside D is related to the area A of D by the formula*

$$\frac{1}{2\pi i} \int_C \phi(z) dz = I - A.$$

Proof: $\int_C \phi(z) dz = \int_C [\zeta(z) - \alpha\bar{z}] dz = \int_C \zeta(z) dz - \alpha \int_C (x - iy)(dx + idy)$. By the Residue Theorem, $(2\pi i)^{-1} \int_C \zeta(z) dz$ is the sum of the residues of $\zeta(z)$ at all interior poles. There are I such poles, and each pole of $\zeta(z)$ has residue 1. Thus $(2\pi i)^{-1} \int_C \zeta(z) dz = I$.

On the other hand, Green's Theorem can be used to show that $\int_C (x - iy)(dx + idy) = \int_C (x - iy) dx + (y + ix) dy = \iint_D (y + ix)_x - (x - iy)_y = \iint_D 2i = 2iA$. Multiply by α and combine with the results of the preceding paragraph to deduce that $\int_C \phi(z) dz = 2\pi iI(D) - 2i\alpha A$. To complete the proof of the lemma, it merely remains to show that $\alpha = \pi$. This can be deduced by taking C to be a square path centered at the origin of sidelength one, encircling the origin once counterclockwise. Since $\phi(z)$ is doubly-periodic, the path integral $\int_C \phi(z) dz$ vanishes. Thus, $0 = 2\pi iI(D) - 2i\alpha A(D)$. For the square, obviously $I = 1$ and $A = 1$. This forces $\alpha = \pi$. Q.E.D.

It is clear from the preceding lemma that the imaginary part of the expression $\phi(z) dz$ provides a link between the number of vertices enclosed by a contour C and the area of the region enclosed by C . We now generalize Lemma 2 by considering curvilinear polygons C that perhaps pass *directly through* finitely many lattice points. It is easy to verify that the singularity of the imaginary part of $\phi(z) dz$ at each lattice point is the same as the singularity of the expression $d\theta = (-ydx + xdy)/(x^2 + y^2)$ at the origin. This singularity is mild enough to give meaning to the improper integral $\int_C \Im[\phi(z) dz]$ whenever C decomposes as the union of finitely-many parametrized paths $C(t)$, along each of which the derivative dC/dt is continuous and *never vanishing*. The improper integral is defined to be the limiting value of the path integral that remains after deletion of the portion of C trapped within a ball of small radius centered at each singularity of the integrand. Since the remaining path of integration is no longer a closed path, we

use the standard trick of introducing an auxiliary circular arc that winds partially around each singularity, that stays within D , and that closes up the disconnected path of integration near the singularity. The resulting closed curve satisfies the hypotheses of Lemma 2. The auxiliary arcs have the effect of introducing correction terms into the conclusion of Lemma 2 that depend upon both the residue of the integrand at each singularity of the integrand and the radian measure of the associated auxiliary arc:

Lemma 3. *Let C be a parametrized simple closed curve as above. Let Θ_j denote the radian measure of the counter-clockwise interior angle within D at the j 'th lattice point on C formed by the incoming and outgoing tangent vectors to C through this vertex (see figure 1). Then,*

$$\frac{1}{2\pi} \int_C \Im[\phi(z) dz] = I - A + \frac{1}{2\pi} \left[\sum_j \Theta_j \right].$$

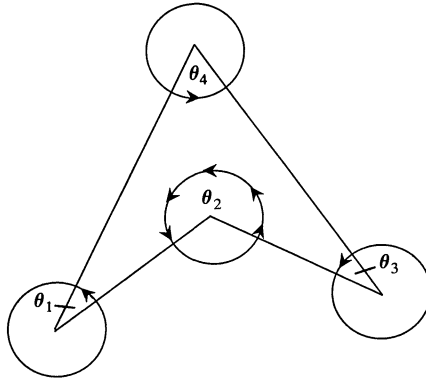


Figure 1

Lemma 4. *For every curve C described above, the path integral of the imaginary part of $\phi(z) dz$ taken along C agrees with the path integral taken along the antipodal curve $\hat{C} = -C$, whose parametrization is related to that of C by $\hat{C}(t) = -C(t)$.*

Proof: It suffices to verify that the expression $\phi(z) dz$ is invariant under the change of variables $z \rightarrow -z$. Recall that ϕ is a linear combination of ζ and $\bar{\zeta}$, both of which are odd functions of z . Thus $\phi(-z)d(-z) = [-\phi(z)](-dz) = \phi(z) dz$. Q.E.D.

Lemma 5. *If the path C in Lemma 4 is actually a polygon, then $\int_C \Im[\phi(z) dz] = 0$.*

Proof: The proof is easier to understand if one first selects a so-called *extreme* vertex of C , which is a vertex with the property that some closed half-plane intersects C only at the vertex. Translate the polygon C so that one such extreme vertex is the origin. Consider the antipodal map $z \rightarrow -z$ that maps C to \hat{C} . The interior of C and the interior of \hat{C} do not overlap, nor do they share any boundary points in common except the extreme vertex at the origin. Apply Lemma 3 to C , and use Lemma 4 to see that $2 \int_C \Im[\phi(z) dz] = \int_C \Im[\phi(z) dz] + \int_{\hat{C}} \Im[\phi(z) dz]$. From

even the crudest diagram it is clear that each line-segment on \hat{C} is a translate of the corresponding antipodal segment on C , *but endowed with the reverse orientation*.¹ The double-periodicity of ϕ therefore implies that the integral of $\Im[\phi(z) dz]$ over $C + \hat{C}$ vanishes because of pairwise cancellation of integrals. Thus $\int_C \Im[\phi(z) dz] = 0$. Q.E.D.

§3. PROOF OF PICK'S FORMULA AND FURTHER COMMENTS. The preceding lemmas provide all the information needed to establish Pick's Formula.

Proof: Apply Lemma 5 to the polygon C . Deduce from Lemma 4 that $A - I = (1/2\pi)\sum_j \Theta_j$. The supplementary angles $\Theta_j - \pi$, sometimes called the clockwise exterior angles to the polygon, sum to -2π because 1 is the total winding number about the origin of the curve swept out by the tangent vector to C [3, p. 217]. Therefore the sum of the interior angles is $\pi(B - 2)$, where B denotes the number of boundary lattice points. Thus $A - I = (1/2)(B - 2)$. Q.E.D.

Comment 1. (Connections with complex cohomology) We had an explicit formula for $\phi(z) dz$ that made it easy to establish Lemmas 1–5. A less explicit proof of the existence of a doubly-periodic function having the properties described in these lemmas could have been established by Hodge-theoretic methods (see [27] for a good introduction to Hodge-Theory). Consider the complex manifold obtained by identifying opposite edges of the unit square to obtain a torus. Place a point mass (δ -function) at the center of the square, and consider the expression $\psi = (2\pi i) [\delta - 1] dx dy$. The integral of this two-form over the unit square vanishes, which by Hodge Theory is the necessary and sufficient condition for there to exist a complex-valued generalized function Φ such that the one-form $\phi(z) dz$ satisfies $d[\Phi(z) dz] = \psi$ on the torus. Note that ψ is invariant under the change of variables $z \rightarrow -z$, since the delta-function and the constant function 1 are even functions with respect to this transformation. To obtain a one-form that is also invariant under this transformation, take $\phi(z)$ to be the odd part of $\Phi(z)$. With this choice we still have $d(\phi dz) = \psi$. Regard ϕ as a doubly-periodic odd function on the complex plane. Lemma 2 now follows upon integration of the identity $d(\phi dz) = \psi$ (Green's Theorem). Lemma 3 follows by taking the imaginary part of Lemma 2 and by noting that the local structure of the singularity of $\phi(z)$ at each lattice point must be like $\zeta(z)$ since $d[\phi dz - \zeta(z) dz]$ is a two-form on the torus that is free of singularities. Lemma 4 follows from the oddness of ϕ . The remaining arguments (Lemma 5 and the proof of Pick's Formula) require no further modifications.

Comment 2. (Connections with magnetostatics) In the theory of classical magnetostatics, the one-form $\Im[\phi(z) dz]$ can be identified with the magnetic force field \vec{B} in R^3 induced by the steady flow of two opposing currents whose combined density distribution is ψ . The singular and regular parts of ψ correspond to a current concentrated on a lattice of parallel wires, and a spatially uniform opposing current, respectively. Lemma 5 can be interpreted as a restricted conservation law satisfied by this induced magnetic field valid for all lattice *polygons* (see figure 2).

¹This is the only part of the proof where we require that the sides of C be line-segments rather than curves.

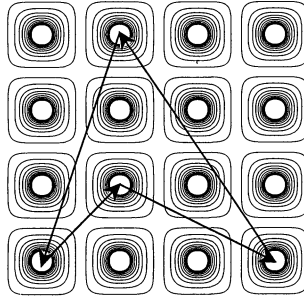


Figure 2

Comment 3. (Connections with hyperbolic lattices) Pick's theorem fails in the hyperbolic plane. That is, there is no linear relationship between the area of a geodesic polygon and the number of lattice points which it encloses. To make this precise, we should first define what a lattice means in the hyperbolic plane. The *Modular Lattice* is defined by

$$\mathcal{L} = \left\{ \frac{a\rho + b}{c\rho + d} \middle| A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL_2(\mathbb{Z}) \right\},$$

where $\rho = e^{2\pi i/3}$. We define a polygon to be a simple closed curve which is the finite union of geodesic arcs (which must be circular arcs perpendicular to the x-axis, or vertical line segments), where the vertices of the polygon are now in \mathcal{L} .

As is well-known [1], Gauss-Bonnet tells us that the area of a hyperbolic triangle is π minus the sum of its interior angles. Fix the triangle $\triangle ABC$, where $A = \rho$, $B = \rho + 1$, $C = i\infty$. Since $\angle ABC = \angle CAB = \pi/3$, and $\angle BCA = 0$, the area of $\triangle ABC$ is $\pi - (\pi/3 + \pi/3 + 0) = \pi/3$. Now fix the triangle $\triangle ABD$, where $D = \rho + 2$. Thus

$$\text{area}(\triangle ABD) = \pi - (\varepsilon + 2\pi/3 + \varepsilon) = \pi/3 - 2\varepsilon,$$

where $\varepsilon > 0$. Thus $\text{area}(\triangle ABD) \neq \text{area}(\triangle ABC)$.

If there were a linear relationship between the area of a geodesic polygon and its boundary and interior lattice points, then both of the above triangles would have the same area, because they both have the same number of interior and boundary lattice points. The fact that they do not have equal areas proves our claim. A similar argument shows that Pick's theorem fails for any lattice in the hyperbolic plane which is generated by a discrete subgroup of $SL_2(\mathbb{R})$. Thus there is something unique about the Archimedean case.

Comment 4. (Connections with the Weierstrass \wp -function) Recall that in the proof of Lemma 2 we established that $\alpha = \pi$. That is, Pick's Formula has provided us with properties of the Weierstrass \wp -function!

REFERENCES

1. A. Beardon, *Geometry of Discrete Groups*, Springer Verlag, 1983.
2. Cappell, S. E. and Shaneson, J. L., Genera of Algebraic Varieties and Counting of Lattice Points, *The Bulletin of the AMS* (Jan. 1994), 62–69.
3. W. R. Derrick, *Complex Analysis and Applications*, 2nd ed., Wadsworth Inc., 1984.
4. E. Ehrhart, Sur un probleme de geometrie diophantienne lineaire I, *J. Reine Angew. Math.* 226 (1967), 25–49.

5. E. Ehrhart, Sur un probleme de geometrie diophantienne lineaire II, *J. Reine Angew. Math.* 227 (1967), 25–49.
6. W. Fulton, *Introduction to Toric Varieties*, Princeton University Press, 1993.
7. W. W. Funkenbusch, From Euler's Formula to Pick's Formula Using an Edge Theorem, *Amer. Math. Monthly* 81 (1974), 647–648.
8. B. Grunbaum and G. C. Shephard, Pick's Theorem, *Amer. Math. Monthly* (1993), 150–161.
9. G. Haigh, A 'natural' approach to Pick's theorem, *The Mathematical Gazette* 64 (1980), 173–177.
10. K. Kolodziejczyk, Areas of Lattice Figures in the Planar Tilings with Congruent Regular Polygons, *J. of Combinatorial Theory, Series A* 58 (1991), 115–126.
11. A. C. F. Liu, Lattice Points and Pick's Theorem, *Math. Magazine* 52 (1979), 232–235.
12. I. G. MacDonald, The volume of a lattice polyhedron, *Proc. Camb. Phil. Soc.* 59 (1963), 719–726.
13. A. I. Markushevich, *Theory of Functions of a Complex Variable*, Chelsea Publishing Company, 1985.
14. P. McMullen, Valuations and Euler-type Relations on certain classes of convex polytopes, *Proc. London Math. Soc.* 35 (1977), 113–135.
15. I. Niven and H. S. Zuckerman, Lattice points and polygonal area, *Amer. Math. Monthly* 74 (1967), 1195–1200.
16. G. Pick, *Geometrisches zur Zahlenlehre*, Sitzungber. Lotos (Prague) 19 (1899), 311–319.
17. C. Polis, Pick's Theorem Extended and Generalized, *Math Teacher* (1991), 399–401.
18. J. Pommersheim, Toric varieties, lattice points, and Dedekind sums, *Math. Annalen* 295, no. 1 (1993), 1.
19. J. E. Reeve, On the volume of lattice polyhedra, *Proc. London Math. Soc.* 7 (1957), 378–395.
20. J. E. Reeve, A further note on the volume of lattice polyhedra, *Proc. London Math. Soc.* 34 (1959), 57–62.
21. D. Ren and J. R. Reay, The Boundary Characteristic and Pick's Theorem in the Archimedean Planar Tilings, *J. of Combinatorial Theory, Series A* (1987), 110–119.
22. B. Reznick, Lattice Point Simplices, *Discrete Math.* 60 (1986), 219–242.
23. P. R. Scott, The Fascination of the Elementary, *Amer. Math. Monthly* 94 (1987), 759–768.
24. R. Stanley, Combinatorial Reciprocity Theorems, *Advances in Math.* 14 (1974), 194–253.
25. H. Steinhaus, *Mathematical Snapshots*, Oxford Univ. Press, New York, 1969.
26. D. E. Varberg, Pick's Theorem Revisited, *Amer. Math. Monthly* 92 (1985), 584–587.
27. Warner, F. W., *Foundations of Differentiable Manifolds and Lie Groups*, Springer Verlag, 1983.

Department of Mathematics
 University of Northern Colorado
 Greeley, CO 80639
 rdiaz@goldng8.univnorthco.edu
 srobins@hopper.univnorthco.edu

A formal manipulator in mathematics often experiences the discomforting feeling that his pencil surpasses him in intelligence.

—Howard W. Eves
In Mathematical Circles. Boston: Prindle, Weber
 and Schmidt, 1969, p. 52.

NOTES

Edited by: John Duncan

Permutations as Products of Transpositions

George Mackiw

When writing a permutation as a product of transpositions, what is the smallest number of transpositions that can be used? This question and variants of it occur both abstractly [2] and in applied settings such as data exchange and sorting [3]. The answer is known and easily stated: the minimum number is precisely $n - r$, where r is the number of disjoint cycles in the given permutation on n letters. One way to establish this result is to use an inductive argument relying on an analysis of how cycles multiply [1]. Another approach, employed in [4], restates the problem in the language of graph theory and makes use of the fact that a connected graph with n vertices must have at least $n - 1$ edges.

The purpose of this note is to provide an alternate derivation of this result that uses only elementary linear algebra. The very answer, $n - r$, seems to suggest some dimension counting involving complementary spaces, and, indeed, our argument takes advantage of orthogonality and the Gram-Schmidt process.

The linear algebraic connection is a natural one. Elements of the symmetric group S_n permute coordinates in \mathbf{R}^n and are often realized as permutation matrices. More precisely, we regard a permutation σ in S_n as acting on the Euclidean space \mathbf{R}^n by $\sigma \mathbf{e}_i = \mathbf{e}_{\sigma(i)}$, where $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$ denotes the natural basis of \mathbf{R}^n .

In this setting, transpositions have a simple geometric interpretation. Given the transposition $\tau = (i, j)$ in S_n , $i < j$, we call the vector $\mathbf{e}_i - \mathbf{e}_j$ in \mathbf{R}^n the vector associated to τ . Notice that τ sends this vector to its negative $\mathbf{e}_j - \mathbf{e}_i$. Further, τ fixes pointwise the collection of $n - 1$ vectors $\{\mathbf{e}_k | k \neq i, j\} \cup \{\mathbf{e}_i + \mathbf{e}_j\}$ which are all orthogonal to $\mathbf{e}_i - \mathbf{e}_j$. Indeed, these $n - 1$ vectors form a basis for the subspace (hyperplane) orthogonal to the vector $\mathbf{e}_i - \mathbf{e}_j$. Simply put, τ acts as the reflection through the hyperplane orthogonal to $\mathbf{e}_i - \mathbf{e}_j$.

We also attach a subspace to any permutation σ —the fixed point space, V_σ , consisting of all vectors x in \mathbf{R}^n with $\sigma x = x$. This is the eigenspace of σ corresponding to the eigenvalue $\lambda = 1$. It always has positive dimension since, for example, any vector all of whose components are equal is in V_σ for any σ . As a matter of fact, it is not difficult to see that the fixed point space is determined by the cycle structure of the permutation. Note that vectors in V_σ must have their i th and j th components agreeing whenever i and j occur in a common cycle of σ . It follows that if σ is written as a product of r disjoint cycles, including trivial cycles containing only one point, then V_σ is r -dimensional with each cycle of σ contributing a basis element in a natural way to V_σ .

For example, the permutation $\sigma = (2, 5, 3)(1, 6)$ in S_7 has a four-dimensional fixed point space in \mathbf{R}^7 that has the vectors $\mathbf{e}_2 + \mathbf{e}_5 + \mathbf{e}_3$, $\mathbf{e}_1 + \mathbf{e}_6$, \mathbf{e}_4 and \mathbf{e}_7 for a basis.

We read products of permutations from right to left. Thus, $(1, 2, 3, 4, 5) = (1, 5)(1, 4)(1, 3)(1, 2)$ expresses a five-cycle as a product of four transpositions. In like fashion, an s -cycle can be written using $s - 1$ transpositions. So, a permutation in S_n consisting of r cycles can be written as a product of $n - r$ transpositions. We are now ready to show that no fewer number of transpositions can be employed. Note that in counting cycles of a permutation we always include trivial one element cycles.

Theorem 1. *A permutation in S_n cannot be written as the product of fewer than $n - r$ transpositions, where r is the number of disjoint cycles in the permutation.*

Proof: Suppose σ in S_n is written as $\sigma = \tau_1\tau_2 \cdots \tau_k$, where the τ_i 's are transpositions. Viewing transpositions as reflections through hyperplanes, let $v_i, i = 1, 2, \dots, k$, be the vectors associated to these transpositions. Recall that v_i is a vector orthogonal to the hyperplane determined by τ_i . The Gram-Schmidt orthogonalization process guarantees the existence of at least $n - k$ linearly independent vectors that are orthogonal to the subspace spanned by the v_i 's. These $n - k$ vectors thus lie in the intersection of the k hyperplanes determined by the transpositions and are thus pointwise fixed by each of the transpositions. Thus these vectors are fixed by σ , and so $\dim V_\sigma \geq n - k$. But, $\dim V_\sigma = r =$ number of cycles in σ . The result $k \geq n - r$ follows. \square

Whenever $\sigma = \tau_1\tau_2 \cdots \tau_k$, a product of transpositions, and k is the minimum number allowed by Theorem 1, we refer to this as a minimal representation of σ . We now use orthogonality to show that a minimal representation must have associated vectors that are linearly independent.

Any σ in S_n determines a direct sum decomposition $\mathbf{R}^n = V_\sigma \oplus V_\sigma^\perp$, where V_σ^\perp denotes the orthogonal complement in \mathbf{R}^n of the fixed point space V_σ . If $\sigma = \tau_1\tau_2 \cdots \tau_k$ is a minimal representation, then $\dim V_\sigma = n - k$. Now the vectors $v_i, i = 1, 2, \dots, k$, associated to the transpositions τ_i are normal vectors to hyperplanes H_i . Since τ_i fixes H_i pointwise, the intersection $\bigcap_{i=1}^k H_i$ is a subspace contained in V_σ . The intersection of these k hyperplanes is the solution space to a k by n homogeneous system of equations, where the i th equation expresses the requirement that a vector in H_i must be orthogonal to v_i . Elementary results concerning rank and solution spaces of systems of equations show that $\dim(\bigcap_{i=1}^k H_i) \geq n - k$, with equality occurring exactly when the normal vectors v_1, v_2, \dots, v_k are linearly independent. We have derived the following result.

Theorem 2. *If the representation $\sigma = \tau_1\tau_2 \cdots \tau_k$ is a minimal one, then the associated vectors v_1, v_2, \dots, v_k are linearly independent and form a basis for V_σ^\perp .* \square

For example, $(1, 6)(3, 4)(4, 6)(1, 3)$ could not be a minimal representation, due to the dependence relation $\mathbf{e}_1 - \mathbf{e}_6 = (\mathbf{e}_3 - \mathbf{e}_4) + (\mathbf{e}_4 - \mathbf{e}_6) + (\mathbf{e}_1 - \mathbf{e}_3)$.

Other reasonably intuitive results about minimal products of transpositions can be obtained using this approach. For example, a minimal representation $\sigma = \tau_1\tau_2 \cdots \tau_k$ must respect the cycle structure of σ . For, suppose that some transposition $\tau_i = (a, b)$ was such that a and b belonged to different cycles of σ . Then the vector $v = \mathbf{e}_a - \mathbf{e}_b$ would not have inner product zero with the vector $w = \sum \mathbf{e}_\alpha$, where α ranges through the elements of the cycle of σ containing a . But this contradicts our result that w is in V_σ , while the associated vector v is in V_σ^\perp . In

particular, no transposition $\tau_i = (a, b)$ can have either a or b belonging to a trivial one element cycle of σ .

The converse of Theorem 2 is also true, though we omit the arguments.

REFERENCES

1. Jozsef Denes, The representation of a permutation as the product of a minimal number of transpositions, and its connection with the theory of graphs, *Publ. Math. Institute Hung. Acad. Sci.* 4 (1959), 63–71.
2. Walter Feit, Roger Lyndon, and Leonard L. Scott, A remark about permutations, *Journal of Combinatorial Theory (A)* 18 (1975) 234–235.
3. Lawrence Fialkow and Hector Salas, Data exchange and permutation length, *Mathematics Magazine* 65 (1992) 188–193.
4. O. P. Lossers, Solution to Problem E3058, *American Mathematical Monthly* 93 (1986), 820–821.

*Dept. of Mathematical Sciences
Loyola College in Maryland
Baltimore, MD 21210
mackiw@loyola.edu*

Congruences Relating the Order of a Group to the Number of Conjugacy Classes

Bjorn Poonen

Let G be a finite group, and let $|G|$ denote its order. Let s be the number of conjugacy classes in G . Burnside, in his 1911 text on the theory of finite groups, used representation theory to prove that if $|G|$ is odd, then $|G| \equiv s \pmod{16}$. (See p. 295 of [1].) On p. 320 of the same book, he left as an exercise to show that if every prime dividing $|G|$ is congruent to 1 modulo 4, then $|G| \equiv s \pmod{32}$. The purpose of this note is to show how elementary counting arguments can yield other congruences in the same spirit. Here is what we will prove:

Theorem. *Let $m \geq 2$ be an integer. If each prime divisor of $|G|$ is congruent to 1 modulo m , then $|G| \equiv s \pmod{2m^2}$.*

Taking $m = 2$ in this theorem yields only $|G| \equiv s \pmod{8}$, which is weaker than Burnside's original result. On the other hand, taking $m = 4$ yields exactly his exercise.

Proof: Let

$$T = \{(g, h) \in G \times G \mid gh \neq hg\}.$$

For each unordered pair $\{C_1, C_2\}$ of cyclic subgroups of G , we may consider the set of (g, h) in $G \times G$ such that the subgroups $\langle g \rangle, \langle h \rangle$ they generate are C_1 and C_2 in some order. Such subsets clearly form a partition of $G \times G$.

Step 1. T is a (disjoint) union of such subsets.

Simply note that

$$(g, h) \notin T \Leftrightarrow \forall x \in \langle g \rangle, \forall y \in \langle h \rangle, xy = yx$$

and the right hand side depends only on the unordered pair $\{\langle g \rangle, \langle h \rangle\}$.

Step 2. Any such subset S lying in T has cardinality a multiple of $2m^2$.

Let $\{C_1, C_2\}$ be the pair of cyclic subgroups corresponding to S , and let n_1, n_2 be their orders. Then

$$S = \{(g_1, g_2) \in C_1 \times C_2 \mid g_i \text{ generates } C_i\} \\ \cup \{(g_2, g_1) \in C_2 \times C_1 \mid g_i \text{ generates } C_i\}.$$

Since C_i has $\phi(n_i)$ generators (where ϕ is Euler's phi function), each of the two sets in the union has size $\phi(n_1)\phi(n_2)$. Moreover, $C_1 \neq C_2$ since $S \subseteq T$, so the union is disjoint, and

$$|S| = 2\phi(n_1)\phi(n_2).$$

Also since $S \subseteq T$, neither C_1 nor C_2 equals $\{1\}$, so we can pick prime divisors p_i of n_i . Then p_i divides $|G|$, so m divides $p_i - 1 = \phi(p_i)$, which divides $\phi(n_i)$. Thus $2m^2$ divides $|S|$.

Step 3. $|T| = |G|(|G| - s)$.

If $g \in G$, let $C_G(g)$ denote the centralizer of g in G , i.e., the subgroup consisting of the h in G which commute with g , and let X_g denote the conjugacy class of $g \in G$. Since X_g is the orbit of g under the conjugation action of G on G , and $C_G(g)$ is the stabilizer of g under this action, we have $|X_g| = |G|/|C_G(g)|$. Now

$$\begin{aligned} |(G \times G) \setminus T| &= \sum_{g \in G} (\text{the number of } h \in G \text{ which commute with } g) \\ &= \sum_{g \in G} |C_G(g)| \\ &= |G| \sum_{g \in G} 1/|X_g| \quad (\text{by the remark above}). \end{aligned}$$

The sum of $1/|X_g|$ over a conjugacy class is 1, so the sum over all of G is s , and

$$\begin{aligned} |(G \times G) \setminus T| &= |G|s \\ |T| &= |G|(|G| - s). \end{aligned}$$

Conclusion of Proof: By Steps 1 and 2, $2m^2$ divides $|T|$, which equals $|G|(|G| - s)$, by Step 3. But $2m^2$ and $|G|$ are relatively prime, since every prime factor of $|G|$ is congruent to 1 modulo m (hence greater than m), and $m \geq 2$. Hence $2m^2$ divides $|G| - s$; i.e. $|G| \equiv s \pmod{2m^2}$, as desired. \square

Question. Can the argument in the $m = 2$ case be modified to obtain Burnside's original result, that if $|G|$ is odd, then $|G| \equiv s \pmod{16}$?

ACKNOWLEDGMENT. This note grew out of an exercise in a course on representation theory given by T. Y. Lam at the University of California at Berkeley. I thank him for teaching an inspiring course and for encouraging me in this work.

REFERENCES

1. W. Burnside, *Theory of Groups of Finite Order* (second edition), Dover, 1911.

*Department of Mathematics
University of California at Berkeley
Berkeley, CA 94720
poonen@math.berkeley.edu*

The Color Invariant for Knots and Links

Peter Andersson

1. INTRODUCTION. A *knot* in three dimensions is a closed curve which can be represented in a diagram with double points where the curve crosses itself transversely. We indicate in the diagram which part of the curve that lies under and which part that lies over the crossing. A system consisting of several curves is called a *link*. *Invariants* are used to show that two curves cannot be deformed to each other. All deformations can be performed with the Reidemeister moves (figure 1) together with deformations not concerning the crossings. For proof see [1].

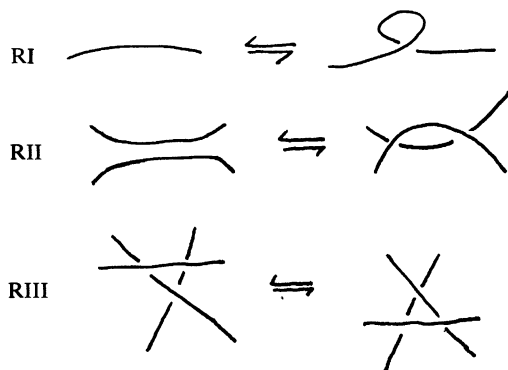


Figure 1. The Reidmeister moves.

2. THE COLOR INVARIANT. An arc is a piece of the curve between two undercrossings. Some other piece of the curve can pass under the arc.

Definition 2.1. A knot or link K can be *colored* mod n if there are integers m_i and n , for each arc in some projection, such that the following holds for all crossings:

$$m_a + m_b \equiv 2m_c \pmod{n} \quad (*)$$

where m_a and m_b are the integers associated with the arcs going under the

crossing, m_c is the integer associated with the overcrossing arc and $n \geq 2$. It is also required that there are colors m_i in at least two different equivalence classes. See [2].

Another formulation of this is that the colors lie symmetrically at the periphery of a circle and that the color of the incoming undercrossing is reflected in the diameter where the color of the overcrossing lies. For $n = 3$ this means that the color at an undercrossing changes if the overcrossing has a different color, and remains the same if it has the same color. We have an odd number of colors at each crossing (Figure 2). See [3] and [4].

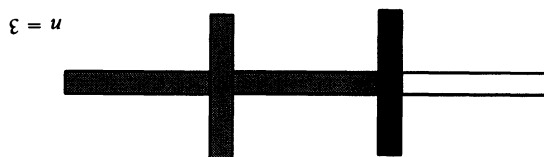


Figure 2.

We have the following well-known theorem (see [2], [3] and [4]):

Theorem 2.2. *Let K be a knot or link which can be colored mod n , then every projection of K can be colored mod n .*

Proof: We check the Reidemeister moves by solving the equations or looking at the circle:

R I: Every arc involved has the same color.

R II: An arc is added or removed. The color of the incoming arc at one undercrossing is reflected back at the other crossing. We can lose one color this way, but it takes at least two different colors to change color at a crossing, so we still have more than one color present.

R III: At most one arc changes color, all colors of the arcs out of the three crossings remain the same. \square

Corollary 2.3. *If a knot can be colored mod n then it cannot be deformed to an unknotted curve.*

Proof: Assume that the knot could be deformed to an unknotted curve. Then by Theorem 2.2 the unknotted curve without crossings could be colored mod n , but a curve with no crossings and only one arc cannot have more than one color. We require that at least two colors are present in a colored projection. This gives a contradiction. \square

Corollary 2.4. *There exists a knot which cannot be deformed to an unknotted curve.*

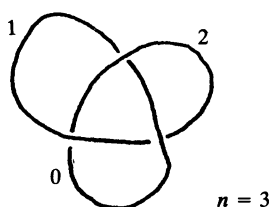


Figure 3. The trefoil knot.

Proof: The trefoil knot can be colored mod 3 (figure 3). By Corollary 2.3 this knot cannot be deformed to an unknotted curve. \square

Corollary 2.5. *If a link is splittable then it can be colored mod n , $n \geq 2$. See Nanyes [2].*

Proof: In a split link the components can be colored with two different colors. By Theorem 2.2 every projection of the link can be colored mod n , $n \geq 2$. \square

Example. The Borromean rings (figure 4). The rings are not pairwise linked. It is easy to see that it is impossible to color the rings mod 3. By Corollary 2.5 this shows that they cannot be split. It is possible, as in the right picture, to color the rings mod 4. We also note that the link is alternating.

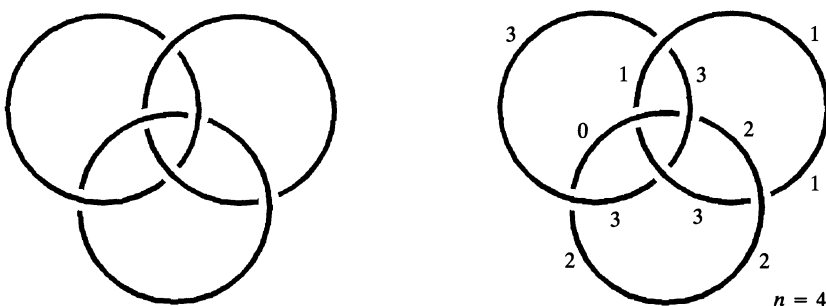


Figure 4. The Borromean rings.

To find if a knot or link K with k crossings can be colored is equivalent to solving a system of linear equations

$$Cx = nb \quad (**)$$

where C is the coefficient matrix corresponding to $n = 0$ in the color relations (*), b and x are integer vectors and n an integer.

Each row in the $k \times k$ matrix C corresponds to a crossing and consists of the elements 2, -1 , -1 and $k - 3$ zeroes. If we as in RI have a loop the row becomes 1, -1 , and zeroes. The columns correspond to the arcs and consist of the elements -1 , -1 , as many twos as the number of curve pieces the arc is crossing over, and zeroes. Addition of the columns gives 0 , so $\det(C) = 0$.

Let C_- be a $k - 1 \times k - 1$ submatrix to C obtained by deleting one row and one column with one non-zero element is common.

To solve (**) we can solve $C_-x = nb_-$ by Cramer's rule. Equations (**) have an integer solution if we choose $n = |\det(C_-)|$. We also need more than one color in the solution to show that the knot or link can be colored.

Example. The knot 8_{20} (figure 5).

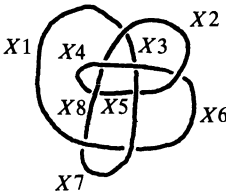


Figure 5. The knot 8_{20} .

With $b = (1, 0, 0, 0, 0, 0, 0, -1)^T$ and $n = |\det(C_-)| = 9$ we see that (**) have a solution $(X1, X2, \dots, X8) = (7, 2, -3, 1, 8, 3, 5, 0)^T$. So the knot can be colored mod 9.

3. ALTERNATING KNOTS AND LINKS

Definition 3.1. A projection of a knot or link is said to be *alternating* if the crossings alternate over-under-over-under-... as one goes along the curve. An alternating projection is reduced if none of the four local regions at a crossing belongs to the same region in the diagram.

If we have a reduced alternating projection the columns of the matrix C consist of $(2, -1, -1, 0, \dots)$ as well as the rows and this holds if we have no simple loops attached to some arc. It is clear that it is possible to write C with the twos as diagonal elements.

Lemma 3.2. Let $\det(C_-)$ be a subdeterminant to the matrix C in (**), written in the form above, with one deleted 2, for a reduced alternating knot or link. Then $\det(C_-)$ is unambiguous.

Proof: If all rows in C_- are summed to a row, the deleted row in C is obtained with different sign. The same holds for the columns. If the new row and column is multiplied by -1 we get the same matrix as if we deleted the other row and column. The determinant is unchanged. \square

Theorem 3.3. Let K be a reduced alternating projection of a knot or link with k crossings. Then K can be colored mod n , for some $n > 1$.

Before we prove the theorem we need to define a matrix.

Definition 3.4. Let C_1^k be a matrix obtained from a submatrix C_-^k to C in (**), for an alternating knot or link with k crossings, by replacing an element in a column with one non-zero element in a row with only two non-zero elements with 1, and the rest of the elements in its column or row are replaced with zeroes. The

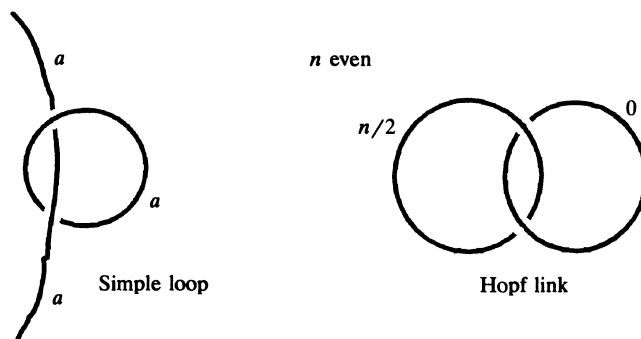


Figure 6.

as the arc and remove it from C . It is easy to find that two simple loops attached to each other (the Hopf link) can be colored mod n , n even. (figure 6) In separated links the columns in the sub-determinants could be linearly dependent and $\det(C_-) = 0$. If K is separated apply Corollary 2.5. \square

Corollary 3.5. *Alternating knots cannot be deformed to an unknotted curve.*

Proof: Follows directly from Theorem 3.3 and Corollary 2.3. \square

This is a new and elementary proof of a well known result which also follows from the behavior of the Jones polynomial. See [3].

Example. The knot 5_1 (Figure 7). C is symmetric and the knot can be colored mod 5.

Example. The figure eight knot with four crossings. (Figure 8)

$$C = \begin{pmatrix} 2 & -1 & 0 & -1 \\ 0 & 2 & -1 & -1 \\ -1 & -1 & 2 & 0 \\ -1 & 0 & -1 & 2 \end{pmatrix}$$

$$|C_-^4| = \begin{vmatrix} 2 & -1 & 0 \\ 0 & 2 & -1 \\ -1 & -1 & 2 \end{vmatrix} = 2 \begin{vmatrix} 2 & -1 \\ -1 & 2 \end{vmatrix} + \begin{vmatrix} 0 & -1 \\ -1 & 2 \end{vmatrix}$$

$$= \begin{vmatrix} 2 & -1 \\ -1 & 2 \end{vmatrix} + \begin{vmatrix} 1 & -1 \\ 0 & 2 \end{vmatrix} = \{|C_-^3| + |C_{1-}^3|\} = 3 + 2 = 5.$$

Take $b_- = (1, 0, 0)^T \Rightarrow X1 = 3, X2 = 1, X3 = 2, X4 = 0.$

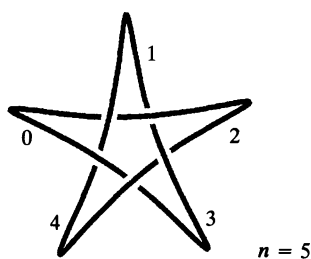


Figure 7. The knot 5_1 .

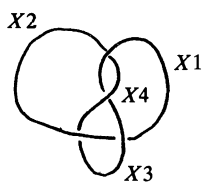


Figure 8. The figure eight knot.

In both examples above the knots can be colored mod 5 and they cannot be unknotted.

4. GENERALIZATIONS OF THE COLOR INVARIANT. The color invariant can be generalized if we orient the curve and the color relations (*) are replaced by:

$$x^*m_a + y^*m_b \equiv (x + y)^*m_c \pmod{n}, \quad (***)$$

where m_a is the integer associated with the outgoing arc in a (+)-crossing or the incoming arc in a (-)-crossing and m_b with the other undergoing arc, m_c is associated with the overcrossing arc and x, y, n are integers. (figure 9)

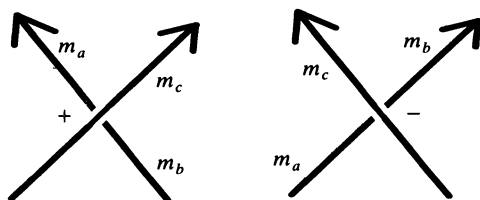


Figure 9. A (+)-crossing to the left and a (-)-crossing to the right.

If n is a prime, i.e., we have a field, n is an invariant as above. The proof of this is a straightforward generalization of the case $x = y = 1$ and is left as an exercise.

Example. With $x = 1$ and $y = 2$ we find that for the figure eight knot $n = 11$ and for the knot 5_1 $n = 31$. So the knots cannot be deformed to each other.

If we put $t = -x/y$ in (***) and calculate $\det(C_-)$, where C_- is a submatrix to C in (**), we formally get the classical Alexander polynomial (see [4], pp. 201–203).

REFERENCES

1. K. Reidemeister. *Knotentheorie*. Chelsea Publishing Co., New York, 1948.
2. O. Nanyes. An elementary proof that the Borromean rings are non-splittable. *Amer. Math. Monthly*, 8 (100) October 1993, 786–789.
3. L. H. Kauffman. New invariants in the theory of knots. *Amer. Math. Monthly*, 3 (95) March 1988, 195–242.
4. L. H. Kauffman. *Knots and Physics*. World Scientific, New Jersey, 1991.

*Rackarbergsgatan 80:210
732 35 Uppsala
Sweden*

THE COMPUTER SCIENCE SAMPLER

Edited by: Catherine C. McGeoch

Veni, Divisi, Vici*

Catherine C. McGeoch

How do you invent a new algorithm for a computational problem? The first source of inspiration is humankind, of course: to develop a sorting algorithm, think about how you would sort a deck of playing cards, or a stack of 200 student papers, and try to write that process down formally. But introspection is not enough. The greatest algorithmic discoveries represent surprising departures from the usual way of doing things.

One of the early landmark events in computer science was Volker Strassen's 1968 discovery that two $n \times n$ matrices could be multiplied using fewer than n^3 scalar multiplications. His algorithm uses $7n^{\log_2 7} - 6n^2$ scalar arithmetic operations where $\log_2 7$ is about 2.808.

Strassen's algorithm is an example of the *divide-and-conquer* paradigm: to solve a problem efficiently, divide it into independent subproblems, recursively solve the subproblems, and recombine the subproblem solutions. Computer scientists have come to recognize about a half-dozen algorithm paradigms, which can guide the search for new algorithms much in the way that Pólya's heuristic strategies (Analogy, Decomposition, Generalization, Induction, etc.) can guide the search for new mathematical results [3]. This column will present Strassen's method and a general technique for analyzing divide-and-conquer algorithms.

THE WAY WE'VE ALWAYS DONE IT. Let X and Y be two $n \times n$ matrices and let Z be their product. The entry in the i th row and j th column of X is denoted x_{ij} , and similarly for Y and Z . The product Z is defined by $z_{ij} = \sum_{k=1}^n x_{ik} \cdot y_{kj}$. The usual method for calculating Z could be written down as follows.

- [1] Set $z_{ij} = 0$ for all pairs (i, j)
- [2] For i ranging from 1 to n , and
- [3] For j ranging from 1 to n , and
- [4] For k ranging from 1 to n , do
- [5] Add $x_{ik} \cdot y_{kj}$ to z_{ij}

To analyze this method we will count the total number of scalar arithmetic operations. Line [5] is performed n^3 times, once for each distinct triple (i, j, k) , and it contains one scalar multiply and one scalar add. Therefore the total number of arithmetic operations is $2n^3$. Note that by changing line [1] to $z_{ij} = x_{i1} \cdot y_{1j}$, we could let k go from 2 to n and save one addition per entry of Z . Then the total number of operations would be $2n^3 - n^2$.

*I came, I divided, I conquered.

DIVIDING WITHOUT CONQUERING. Before describing Strassen's method we will develop a simpler divide-and-conquer strategy. Actually this strategy should be called divide-and-be-conquered because it will turn out to be worse than the classic method; but it will serve to introduce some useful notation and the analysis technique.

Assume for now that n is a power of 2, so $n = 2^k$ for some $k \geq 0$. The $n \times n$ matrix Z can be divided into four $n/2 \times n/2$ matrices Z_{11} , Z_{12} , Z_{21} and Z_{22} located in the upper left, upper right, lower left, and lower right corners respectively. We will call these smaller matrices the *quadrants* of Z . With X and Y divided similarly, it holds that

$$\begin{aligned} Z_{11} &= X_{11} \cdot Y_{11} + X_{12} \cdot Y_{21} \\ Z_{12} &= X_{11} \cdot Y_{12} + X_{12} \cdot Y_{22} \\ Z_{21} &= X_{21} \cdot Y_{11} + X_{22} \cdot Y_{21} \\ Z_{22} &= X_{21} \cdot Y_{12} + X_{22} \cdot Y_{22}. \end{aligned} \tag{1}$$

Thus for n a power of 2 we can obtain the product of X and Y recursively, using eight multiplications and four additions of quadrant matrices. When $n = 1$ the single entry of Z is computed by scalar multiplication. This recursive definition suggests a procedure for matrix multiplication that is sketched below.

- [0] Procedure **Multiply** (X, Y, Z, n) calculates Z as follows:
- [1] If $n = 1$ then let $z_{11} = x_{11} \cdot y_{11}$ and exit the procedure.
- [2] Otherwise, do the following:
- [3] Divide X into quadrants $X_{11}, X_{12}, X_{21}, X_{22}$.
- [4] Divide Y into quadrants $Y_{11}, Y_{12}, Y_{21}, Y_{22}$.
- [5] Apply **Multiply** ($X_{11}, Y_{11}, T_1, n/2$) with result T_1 .
- [6] Apply **Multiply** ($X_{12}, Y_{21}, T_2, n/2$) with result T_2 .
- [7..12] Apply **Multiply** six more times with results $T_3 \dots T_8$.
- [13] Add pairs of matrices from $T_1 \dots T_8$ as in (1) to obtain Z_{11}, Z_{12}, Z_{21} and Z_{22} .
- [14] Combine quadrants to form Z .
- [15] Exit the procedure.

This definition deserves some explanation for those unfamiliar with computer programming. The **Multiply** procedure can be applied to matrices X, Y , and Z of arbitrary size $n = 2^k$; therefore we write it down in terms of parameters (X, Y, Z, n). The general multiplication process involves several separate *instantiations*, or copies, of the procedure. For a given instantiation, the **Apply** operation on Line [5] creates a new instantiation of **Multiply** and sets up a correspondence between $(X_{11}, Y_{11}, T_1, n/2)$, and the parameters (X, Y, Z, n) , respectively. The new instantiation performs **Multiply** on the smaller matrices, producing T_1 . Another instantiation is created on line [6] to calculate T_2 , and so forth.

To analyze this algorithm we use a recurrence formula $T(n)$ that describes the total number of scalar operations required to multiply two $n \times n$ matrices. When $n = 1$ we have $T(1) = 1$. For $n > 1$, line [13] performs four matrix additions on $n/2 \times n/2$ matrices, for a total of n^2 scalar additions. The **Apply** operation occurs eight times, each time instantiating a procedure for $n/2 \times n/2$ matrices. Therefore the recurrence is given by

$$\begin{aligned} T(1) &= 1 \\ T(n) &= 8T(n/2) + n^2 \quad \text{for } n = 2^k, k \geq 1. \end{aligned}$$

Now let's derive a solution to the above recurrence. We will use the facts that $T(1) = 1$ and $T(2) = 12$. It is convenient to work with a new formula defined by $t(k) = T(2^k)$; therefore we have

$$t(k) = 8t(k-1) + 4^k \quad \text{for } k \geq 1.$$

We need to solve

$$t(k) - 8t(k-1) = 4^k, \quad (2)$$

subject to $t(0) = 1$, $t(1) = 12$. Multiply throughout by 4 and substitute $k-1$ for k to obtain

$$4t(k-1) - 32t(k-2) = 4^k \quad \text{for } k > 1.$$

Subtracting this from (2), we have

$$t(k) - 12t(k-1) + 32t(k-2) = 0 \quad \text{for } k > 1. \quad (3)$$

We can solve this using a *characteristic equation* that maps the coefficients of the recurrence into a polynomial in x :

$$x^2 - 12x + 32 = 0.$$

This polynomial has roots $x = 8$ and $x = 4$. It is easy to verify that the following formula satisfies (3):

$$t(k) = c_1(8^k) + c_2(4^k) \quad \text{for } k > 1.$$

The coefficients c_1 and c_2 are determined by initial conditions: here we have $t(0) = 1 = c_1 + c_2$ and $t(1) = 12 = 8c_1 + 4c_2$, which produces $c_1 = 2$ and $c_2 = -1$. Returning to the original notation we have

$$t(k) = 2 \cdot 8^k - 4^k \quad \text{for } k \geq 0$$

$$T(n) = 2 \cdot 8^{\log_2 n} - 4^{\log_2 n} \quad \text{for } n = 2^k, k \geq 0$$

$$T(n) = 2n^3 - n^2 \quad \text{for } n = 2^k, k \geq 0.$$

When n is a power of two this recursive procedure requires exactly the same number of scalar operations as the standard method. If n is not a power of two we can find the smallest $m = 2^k$ such that $n < m$, and imbed X and Y in larger matrices of size $m \times m$, padding with 0's as needed. In this case the recursive procedure is *worse* than the original method!

Strassen's Algorithm. Strassen's matrix multiplication algorithm is similar to the recursive strategy above, but it does not calculate the intermediate terms $T_1 \dots T_8$. Instead it finds different intermediate terms $M_1 \dots M_7$, defined by

$$M_1 = (X_{12} - X_{22}) \cdot (Y_{21} + Y_{22})$$

$$M_2 = (X_{11} + X_{22}) \cdot (Y_{11} + Y_{22})$$

$$M_3 = (X_{11} - X_{21}) \cdot (Y_{11} + Y_{12})$$

$$M_4 = (X_{11} + X_{12}) \cdot Y_{22}$$

$$M_5 = X_{11} \cdot (Y_{12} - Y_{22})$$

$$M_6 = X_{22} \cdot (Y_{21} - Y_{11})$$

$$M_7 = (X_{21} + X_{22}) \cdot Y_{11}.$$

It is straightforward to verify that

$$Z_{11} = M_1 + M_2 - M_4 + M_6$$

$$Z_{12} = M_4 + M_5$$

$$Z_{21} = M_6 + M_7$$

$$Z_{22} = M_2 - M_3 + M_5 - M_7.$$

Let $S(n)$ denote the total number of scalar arithmetic operations performed by Strassen's algorithm. We have $S(1) = 1$ as before. A given instantiation with parameter n performs eighteen matrix additions and subtractions on quadrants, for a total of $18n^2/4 = 9n^2/2$ scalar operations. Also, seven **Apply** steps are needed to perform matrix multiplication on quadrants. Therefore we have $S(n) = 7S(n/2) + 9n^2/2$.

The key observation here is that Strassen's algorithm performs only seven matrix multiplications per instantiation rather than eight. Proceeding with the analysis, we must solve the relation $s(k) - 7s(k-1) = (9/2)4^k$ or equivalently $s(k) - 11s(k-1) + 28s(k-2) = 0$. With the recurrence in this form, we can apply the characteristic equation $x^2 - 11x + 28 = 0$ which has roots $x = 7$ and $x = 4$. Therefore we have $s(k) = c_1 7^k + c_2 4^k$, with coefficients determined by initial conditions $s(0) = 1$ and $s(1) = 25$. The result is $s(k) = 7 \cdot 7^k - 6 \cdot 4^k$, so we obtain

$$S(n) = 7 \cdot 7^{\log_2 n} - 6 \cdot 4^{\log_2 n}$$

$$S(n) = 7n^{\log_2 7} - 6n^2, \text{ for } n = 2^k, k \geq 0$$

where $\log_2 7 \approx 2.808$.

The smallest n for which $7n^{2.808} - 6n^2 \geq 2n^3 - n^2$ holds is 668. If you ever need to multiply two 1024×1024 matrices you could save about 166 million scalar operations (with only about 1.98 billion remaining) by using Strassen's algorithm instead of the usual method. Although the algorithm can be implemented so that no actual costs are incurred in lines [3], [4], and [14], the costs of instantiating new procedures and of handling cases where n is not a power of 2 combine to make Strassen's algorithm more of theoretical than practical interest.

Strassen's discovery prompted an intensive worldwide search for even better matrix multiplication algorithms. The first improvement came ten years later, when Victor Pan showed that the leading exponent could be lowered to 2.795 (see [2] for a discussion of research on this problem up to 1982). Currently the best known algorithm has an exponent of 2.376 [1]. It is known that the exponent must be least 2; obtaining either a better multiplication algorithm or a higher lower bound remains one of the most famous open problems in algorithm analysis.

Addendum. Try to invent a divide-and-conquer algorithm for sorting n numbers. At least two such algorithms are known; you can read about Quicksort and Mergesort in any textbook on algorithms.

ACKNOWLEDGMENT. I thank David Armacost for his help with the title.

REFERENCES

1. D. Coppersmith and S. Winograd, Matrix multiplication via arithmetic progressions, *Journal of Symbolic Computation* 9, 1990, pp 251–280.
2. V. Pan, *How to Multiply Matrices Faster*, Lecture Notes in Computer Science No 179, Springer-Verlag, 1982.
3. G. Pólya *How to Solve It*, Doubleday, 1957.
4. V. Strassen, Gaussian Elimination is not optimal, *Numerische Mathematik*, 13 (1969) 354–356.

Department of Mathematics
Amherst College
Amherst, MA 01002
ccm@cs.amherst.edu

THE EVOLUTION OF . . .

Edited by **Abe Shenitzer**

Mathematics, York University, North York, Ontario M3J 1P3, Canada

Part I. Topology and Abstract Algebra as Two Roads of Mathematical Comprehension*

Unterrichtsblätter für Mathematik und Naturwissenschaften 38, 177–188 (1932). (A lecture in the summer course of the Swiss Society of Gymnasium Teachers, given in Bern, in October 1931.)

Hermann Weyl

We are not very pleased when we are forced to accept a mathematical truth by virtue of a complicated chain of formal conclusions and computations, which we traverse blindly, link by link, feeling our way by touch. We want first an overview of the aim and of the road; we want to understand the *idea* of the proof, the deeper context. A modern mathematical proof is not very different from a modern machine, or a modern test setup: the simple fundamental principles are hidden and almost invisible under a mass of technical details. When discussing Riemann in his lectures on the history of mathematics in the 19th century, Felix Klein said:

Undoubtedly, the capstone of every mathematical theory is a convincing proof of all of its assertions. Undoubtedly, mathematics inculcates itself when it foregoes convincing proofs. But the mystery of brilliant productivity will always be the posing of new questions, the anticipation of new theorems that make accessible valuable results and connections. Without the creation of new viewpoints, without the statement of new aims, mathematics would soon exhaust itself in the rigor of its logical proofs and begin to stagnate as its substance vanishes. Thus, in a sense, mathematics has been most advanced by those who distinguished themselves by intuition rather than by rigorous proofs.

The key element of Klein's own method was an intuitive perception of inner connections and relations whose foundations are scattered. To some extent, he failed when it came to a concentrated and pointed logical effort. In his commemorative address for Dirichlet, Minkowski contrasted the minimum principle that Germans tend to name for Dirichlet (and that was actually applied most comprehensively by William Thomson) with the true Dirichlet principle: to conquer problems with a minimum of blind computation and a maximum of insightful thoughts. It was Dirichlet, said Minkowski, who ushered in the new era in the history of mathematics.

What is the secret of such an understanding of mathematical matters, what does it consist in? Recently, there have been attempts in the philosophy of science to contrast understanding, the art of interpretation as the basis of the humanities, with scientific explanation, and the words intuition and understanding have been

*The original German version of this article is found in vol. 3, pp. 348–358, of the four-volume edition of Hermann Weyl's collected works published by Springer-Verlag in 1968. The translation is by Abe Shenitzer.

invested in this philosophy with a certain mystical halo, an intrinsic depth and immediacy. In mathematics, we prefer to look at things somewhat more soberly. I cannot enter into these matters here, and it strikes me as very difficult to give a precise analysis of the relevant mental acts. But at least I can single out, from the many characteristics of the process of understanding, one that is of decisive importance. One separates in a natural way the different aspects of a subject of mathematical investigation, makes each accessible through its own relatively narrow and easily surveyable group of assumptions, and returns to the complex whole by combining the appropriately specialized partial results. This last synthetic step is purely mechanical. The great art is in the first, analytic, step of appropriate separation and generalization. The mathematics of the last few decades has revelled in generalizations and formalizations. But to think that mathematics pursues generality for the sake of generality is to misunderstand the sound truth that a natural generalization *simplifies* by reducing the number of assumptions and by thus letting us understand certain aspects of a disarranged whole. Of course, it can happen that different directions of generalization enable us to understand different aspects of a particular concrete issue. Then it is subjective and dogmatic arbitrariness to speak of the true ground, the true source of an issue. Perhaps the only criterion of the naturalness of a severance and an associated generalization is their fruitfulness. If this process is systematized according to subject matter by a researcher with a measure of skill and “sensitive fingertips” who relies on all the analogies derived from his experience, then we arrive at axiomatics, which today is an instrument of concrete mathematical investigation rather than a method for the clarification and “deep-laying” of foundations.

In recent years mathematicians have had to focus on the general and on formalization to such an extent that, predictably, there have turned up many instances of cheap and easy generalizing for its own sake. Pólya has called it generalizing by dilution. It does not increase the essential mathematical substance. It is much like stretching a meal by thinning the soup. It is deterioration rather than improvement. The aged Klein said: “Mathematics looks to me like a store that sells weapons in peacetime. Its windows are replete with luxury items whose ingenious, artful and eye-catching execution delights the connoisseur. The true origin and purpose of these objects—the strike that defeats the enemy—have receded into the background and have been all but forgotten.” There is perhaps more than a grain of truth in this indictment, but, on the whole, our generation regards this evaluation of its efforts as unjust.

There are two modes of understanding that have proved, in our time, to be especially penetrating and fruitful. The two are topology and abstract algebra. A large part of mathematics bears the imprint of these two modes of thought. What this is attributable to can be made plausible at the outset by considering the central concept of real number. The system of real numbers is like a Janus head with two oppositely directed faces. In one respect it is the domain of the operations $+$ and \times and their inverses, in another it is a continuous manifold, and the two are continuously related. One is the algebraic and the other is the topological face of numbers. Since modern axiomatics is simpleminded and (unlike modern politics) dislikes such ambiguous mixtures of peace and war, it made a clean break between the two. The notion of size of number, expressed in the relations $<$ and $>$, occupies a kind of intermediate relation between algebra and topology.

Investigations of continua are purely topological if they are restricted to just those properties and differences that are unchanged by arbitrary continuous

deformations, by arbitrary continuous mappings. The mappings in question need only be faithful to the extent to which they don't collapse what is distinct. Thus it is a topological property of a surface to be closed like the surface of a sphere or open like the ordinary plane. A piece of the plane is said to be simply connected if, like the interior of a circle, it is partitioned by every crosscut. On the other hand, an annulus is doubly connected because there exists a crosscut that does not partition it but every subsequent crosscut does. Every closed curve on the surface of a sphere can be shrunk to a point by means of a continuous deformation, but this is not the case for a torus. Two closed curves in space can be intertwined or not. These are examples of topological properties or dispositions. They involve the primitive differences that underlie all finer differentiations of geometric figures. They are based on the single idea of continuous connection. References to a particular structure of a continuous manifold, such as a metric, are foreign to them. Other relevant concepts are limit, convergence of a sequence of points to a point, neighborhood and continuous line.

After this preliminary sketch of topology I want to tell you briefly about the motives that have led to the development of abstract algebra. Then I will use a simple example to show how the same issue can be looked at from a topological and from an abstract-algebraic viewpoint.

All a pure algebraist can do with numbers is apply to them the four operations of addition, subtraction, multiplication and division. If a system of numbers is a field, that is, if it is closed under these operations, then the algebraist has no means of going beyond it. The simplest field is the field of rationals. Another example is the field of numbers of the form $a + b\sqrt{2}$, a, b rational. The well-known concept of irreducibility of polynomials is relative and depends on the field of coefficients of the polynomials, namely a polynomial $f(x)$ with coefficients in a field K is said to be irreducible over K if it cannot be written as a product $f_1(x) \cdot f_2(x)$ of two non-constant polynomials with coefficients in K . The solution of linear equations and the determination of the greatest common divisor of two polynomials by means of the Euclidean algorithm are carried out within the field of the coefficients of the equations and of the polynomials respectively. The classical problem of algebra is the solution of an algebraic equation $f(x) = 0$ with coefficients in a field K , say the field of rationals. If we know a root ϑ of the equation, then we know the numbers obtained by applying to ϑ and to the (presumably known) numbers in K the four algebraic operations. The resulting numbers form a field $K(\vartheta)$ that contains K . In $K(\vartheta)$, ϑ plays a role of a determining number from which all other numbers in $K(\vartheta)$ are rationally derivable. But many—virtually all—numbers in $K(\vartheta)$ can play the same role as ϑ . It is therefore a breakthrough if we replace the study of the equation $f(x) = 0$ by the study of the field $K(\vartheta)$. By doing this we eliminate all manner of trivia and consider at the same time all equations that can be obtained from $f(x) = 0$ by means of Tschirnhausen transformations. The algebraic, and above all the arithmetical, theory of number fields is one of the sublime creations of mathematics. From the viewpoint of the richness and depth of its results it is the most perfect such creation.

There are fields in algebra whose elements are not numbers. The polynomials in one variable, or indeterminate, x , [with coefficients in a field], are closed under addition, subtraction and multiplication but not under division. Such a system of magnitudes is called an integral domain. The idea that the argument x is a variable that traverses continuously its values is foreign to algebra; it is just an indeterminate, an empty symbol that binds the coefficients of the polynomial into a

uniform expression that makes it easier to remember the rules for addition and multiplication. 0 is the polynomial all of whose coefficients are 0 (not the polynomial which takes on the value 0 for all values of the variable x). It can be shown that the product of two nonzero polynomials is $\neq 0$. The algebraic viewpoint does not rule out the substitution for x of a number a taken from the field in which we operate. But we can also substitute for x a polynomial in one or more indeterminates y, z, \dots . Such substitution is a formal process which effects a faithful projection of the integral domain $K[x]$ of polynomials in x onto K or onto the integral domain of polynomials $K[y, z, \dots]$; here “faithful” means subject to the preservation of the relations established by addition and multiplication. It is this formal operating with polynomials that we are required to teach students studying algebra in school. If we form quotients of polynomials, then we obtain a field of rational functions which must be treated in the same formal manner. This, then, is a field whose elements are functions rather than numbers. Similarly, the polynomials and rational functions in two or three variables, x, y or x, y, z with coefficients in K form an integral domain and field respectively.

Compare the following three integral domains: the integers, the polynomials in x with rational coefficients, and the polynomials in x and y with rational coefficients. The Euclidean algorithm holds in the first two of these domains, and so we have the theorem: If a, b are two relatively prime elements, then there are elements p, q in the appropriate domain such that

$$(*) \quad 1 = p \cdot a + q \cdot b.$$

This implies that the two domains in question are unique factorization domains. The theorem $(*)$ fails for polynomials in two variables. For example, $x - y$ and $x + y$ are relatively prime polynomials such that for every choice of polynomials $p(x, y)$ and $q(x, y)$ the constant term of the polynomial $p(x, y)(x - y) + q(x, y)(x + y)$ is 0 rather than 1. Nevertheless polynomials in two variables with coefficients in a field form a unique factorization domain. This example points to interesting similarities and differences.

There is yet another way of making fields in algebra. It involves neither numbers nor functions but congruences. Let p be a prime integer. Identify two integers if their difference is divisible by p , or, briefly, if they are congruent mod p . (To “see” what this means wrap the real line around a circle of circumference p .) The result is a field with p elements. This representation is extremely useful in all of number theory. Consider, for example, the following theorem of Gauss that has numerous applications: If $f(x)$ and $g(x)$ are two polynomials with integer coefficients such that all coefficients of the product $f(x) \cdot g(x)$ are divisible by a prime p , then all coefficients of $f(x)$ or all coefficients of $g(x)$ are divisible by p . This is just the trivial theorem that the product of two polynomials can be 0 only if one of its factors is 0, applied to the field just described as the field of coefficients. This integral domain contains polynomials that are not 0 but vanish for all values of the argument; one such polynomial is $x^p - x$. In fact, by Fermat’s theorem, we have

$$a^p - a \equiv 0 \pmod{p}.$$

Cauchy uses a similar approach to construct the complex numbers. He regards the imaginary unit i as an indeterminate and studies polynomials in i over the reals modulo $i^2 + 1$, that is he regards two polynomials as equal if their difference is divisible by $i^2 + 1$. In this way, the actually unsolvable equation $i^2 + 1 = 0$ is rendered, in some measure, solvable. Note that the polynomial $i^2 + 1$ is prime over the reals. Kronecker generalized Cauchy’s construction as follows. Let K be a field and $p(x)$ a polynomial prime over K . Viewed modulo $p(x)$, the polynomials

$f(x)$ with coefficients in K form a field (and not just an integral domain). From an algebraic viewpoint, this process is fully equivalent to the one described previously, and can be thought of as the process of extending K to $K(\vartheta)$ by adjoining to K a root of the equation $p(\vartheta) = 0$. But it has the advantage that it takes place within pure algebra and gets around the demand for solving an equation that is actually unsolvable over K .

It is quite natural that these developments should have prompted a purely axiomatic buildup of algebra. A field is a system of objects, called numbers, closed under two operations, called addition and multiplication, that satisfy the usual axioms: both operations are associative and commutative, multiplication is distributive over addition, and both operations are uniquely invertible yielding subtraction and division respectively. If the axiom of invertibility of multiplication is left out, then the resulting system is called a ring. Now “field” no longer denotes, as before, a kind of sector of the continuum of real or complex numbers but a self-contained universe. One can apply the field operations to elements of the same field but not to elements of different fields. In this process we need not resort to artificial abstracting from the size relations $<$ and $>$. These relations are irrelevant for algebra and the “numbers” of an abstract “number field” are not subject to such relations. In place of the uniform number continuum of analysis we now have the infinite multiplicity of structurally different fields. The previously described processes, namely adjunction of an indeterminate and identification of elements that are congruent with respect to a fixed prime element, are now seen as two modes of construction that lead from rings and fields to other rings and fields respectively.

The elementary axiomatic grounding of geometry also leads to this abstract number concept. Take the case of plane projective geometry. The incidence axioms alone lead to a “number field” that is naturally associated with it. Its elements, the “numbers,” are purely geometric entities, namely dilations. A point and a straight line are ratios of triples of “numbers” in that field, $x_1 : x_2 : x_3$ and $u_1 : u_2 : u_3$ respectively, such that incidence of the point $x_1 : x_2 : x_3$ on the line $u_1 : u_2 : u_3$ is represented by the equation

$$x_1 u_1 + x_2 u_2 + x_3 u_3 = 0.$$

Conversely, if one uses these algebraic expressions to define the geometric terms, then every abstract field leads to an associated projective plane that satisfies the incidence axioms. It follows that a restriction involving the number field associated with the projective plane cannot be read off from the incidence axioms. Here the preexisting harmony between geometry and algebra comes to light in the most impressive manner. For the geometric number system to coincide with the continuum of ordinary real numbers one must introduce axioms of order and continuity, very different in kind from the incidence axioms. We thus arrive at a reversal of the development that has dominated mathematics for centuries and seems to have arisen originally in India and to have been transmitted to the West by Arab scholars: Up till now, we have regarded the number concept as the logical antecedent of geometry, and have therefore approached every realm of magnitudes with a universal and systematically developed number concept independent of the applications involved. Now, however, we revert to the Greek viewpoint that every subject has an associated intrinsic number realm that must be derived from within it. We experience this reversal not only in geometry but also in the new quantum physics. According to quantum physics, the physical magnitudes associated with a particular physical setup (*not* the numerical values that they may take

on depending on its different states) admit of an addition and a non-commutative multiplication, and thus give rise to a system of algebraic magnitudes intrinsic to it that cannot be viewed as a sector of the system of real numbers.

And now, as promised, I will present a simple example that illustrates the mutual relation between the topological and abstract-algebraic modes of analysis. I consider the theory of algebraic functions of a single variable x . Let $K(x)$ be the field of rational functions of x with arbitrary complex coefficients. Let $f(z)$, more precisely $f(z; x)$, be an n -th degree polynomial in z with coefficients in $K(x)$. We explained earlier when such a polynomial is said to be irreducible over $K(x)$. This is a purely algebraic concept. Now construct the Riemann surface of the n -valued algebraic function $z(x)$ determined by the equation $f(z; x) = 0$. Its n sheets extend over the x -plane. For easier transformation of the x -plane into the x -sphere by means of a stereographic projection we add to the x -plane a point at infinity. Like the sphere, our Riemann surface is now closed. The irreducibility of the polynomial f is reflected in a very simple topological property of the Riemann surface of $z(x)$, namely its connectedness: if we shake a paper model of that surface it does not break into distinct pieces. Here you witness the coincidence of a purely algebraic and a purely topological concept. Each suggests generalization in a different direction. The algebraic concept of irreducibility depends only on the fact that the coefficients of the polynomial are in a field. In particular, $K(x)$ can be replaced by the field of rational functions of x with coefficients in a preassigned field k which takes the place of the continuum of all complex numbers. On the other hand, from the viewpoint of topology it is irrelevant that the surface in question is a Riemann surface, that it is equipped with a conformal structure, and that it consists of a finite number of sheets that extend over the x -plane. Each of the two antagonists can accuse the other of admitting side issues and of neglecting essential features. Who is right? Questions such as these, involving not facts but ways of looking at facts, can lead to hatred and bloodshed when they touch human emotions. In mathematics, the consequences are not so serious. Nevertheless, the contrast between Riemann's topological theory of algebraic functions and Weierstrass' more algebraically directed school led to a split in the ranks of mathematicians that lasted for almost a generation.

Weierstrass himself wrote to his faithful pupil H. A. Schwarz: "The more I reflect on the principles of function theory—and I do this all the time—the stronger is my conviction that this theory must be established on the foundation of algebraic truths, and that it is therefore not the right way when, contrariwise, the 'transcendent' (to put it briefly) is invoked to establish simple and fundamental algebraic theorems—this is so no matter how attractive are, at a first glance, say, the considerations by means of which Riemann discovered so many of the most important properties of algebraic functions." This strikes us now as onesided; neither one of the two ways of understanding, the topological or the algebraic, can be acknowledged to have unconditional advantage over the other. And we cannot spare Weierstrass the reproach that he stopped midway. True, he explicitly constructed the functions as algebraic, but he also used as coefficients the algebraically unanalyzed, and in a sense unfathomable for algebraists, continuum of complex numbers. The dominant general theory in the direction followed by Weierstrass is the theory of an abstract number field and its extensions determined by means of algebraic equations. Then the theory of algebraic functions moves in the direction of a shared axiomatic basis with the theory of algebraic numbers. In fact, what suggested to Hilbert his approaches in the theory of number fields was the analogy [between the latter] and the state of things in the realm of algebraic

functions discovered by Riemann by his topological methods. (Of course, when it came to proofs, the analogy was useless.)

Our example “irreducible-connected” is typical also in another respect. How visually simple and understandable is the topological criterion (shake the paper model and see if it falls apart) in comparison with the algebraic! The visual primality of the continuum (I think that in this respect it is superior to the 1 and the natural numbers) makes the topological method particularly suitable for both discovery and synopsis in mathematical areas, but is also the cause of difficulties when it comes to rigorous proofs. While it is close to the visual, it is also refractory to logical approaches. That is why Weierstrass, M. Noether and others preferred the laborious, but more solid-feeling, procedure of direct algebraic construction to Riemann’s transcendental-topological justification. Now, step by step, abstract algebra tidies up the clumsy computational apparatus. The generality of the assumptions and axiomatization force one to abandon the path of blind computation and to break the complex state of affairs into simple parts that can be handled by means of simple reasoning. Thus algebra turns out to be the El Dorado of axiomatics.

I must add a few words about the method of topology to prevent the picture from becoming altogether vague. If a continuum, say, a two-dimensional closed manifold, a surface, is to be the subject of mathematical investigation, then we must think of it as being subdivided into finitely many “elementary pieces” whose topological nature is that of a circular disk. These pieces are further fragmented by repeated subdivision in accordance with a fixed scheme, and thus a particular spot in the continuum is ever more precisely intercepted by an infinite sequence of nested fragments that arise in the course of successive subdivisions. In the one-dimensional case, the repeated “normal subdivision” of an elementary segment is its bipartition. In the two-dimensional case, each edge is first bipartitioned, then each piece of surface is divided into triangles by means of lines in the surface that lead from an arbitrary center to the (old and new) vertices. What proves that a piece is elementary is that it can be broken into arbitrarily small pieces by repetition of this division process. The scheme of the initial subdivision into elementary pieces—to be referred to briefly in what follows as the “skeleton”—is best described by labelling the surface pieces, edges and vertices by means of symbols, and thus prescribing the mutual bounding relations of these elements. Following the successive subdivisions, the manifold may be said to be spanned by an increasingly dense net of coordinates which makes it possible to determine a particular point by means of an infinite sequence of symbols that play a role comparable to that of numbers. The reals appear here in the particular form of dyadic fractions, and serve to describe the subdivision of an open one-dimensional continuum. Other than that, we can say that each continuum has its own arithmetical scheme; the introduction of numerical coordinates by reference to the special division scheme of an open one-dimensional continuum violates the nature of things, and its sole justification is the practical one of the extraordinary convenience of the calculational manipulation of the continuum of numbers with its four operations. In the case of an actual continuum, the subdivisions can be realized only with a measure of imprecision; one must imagine that, as the process of subdivision progresses step by step, the boundaries set by the earlier subdivisions are ever more sharply fixed. Also, in the case of an actual continuum, the process of subdivision that runs virtually ad infinitum can reach only a certain definite stage. But in distinction to concrete realization, the localization in an actual continuum, the combinatorial scheme, the arithmetical nullform, is a priori deter-

mined *ad infinitum*; and mathematics deals with this combinatorial scheme alone. Since the continued subdivision of the initial topological skeleton progresses in accordance with a fixed scheme, it must be possible to read off all the topological properties of the nascent manifold from that skeleton. This means that, in principle, it must be possible to pursue topology as finite combinatorics. For topology, the ultimate elements, the atoms, are, in a sense, the elementary parts of the skeleton and not the points of the relevant continuous manifold. In particular, given two such skeletons, it must be possible to decide if they lead to concurrent manifolds. Put differently, it must be possible to decide if we can view them as subdivisions of one and the same manifold.

The algebraic counterpart of the transition from the algebraic equation $f(z; x) = 0$ to the Riemann surface is the transition from the latter equation to the field determined by the function $z(x)$; this is so because the Riemann surface is uniquely occupied not only by the function $z(x)$ but also by all algebraic functions in this field. What is characteristic for Riemann's function theory is the converse problem: given a Riemann surface construct its field of algebraic functions. The problem has always just one solution. Since every point \wp of the Riemann surface lies over a definite point of the x -plane, the Riemann surface, as presently constituted, is embedded in the x -plane. The next step is to abstract from the embedding relation $\wp \rightarrow x$. As a result, the Riemann surface becomes, so to say, a free-floating surface equipped with a conformal structure and an angle measure. Note that in ordinary surface theory we must learn to distinguish between the surface as a continuous structure made up of elements of a specific kind, its points, and the embedding in 3-space that associates with each point \wp of the surface, in a continuous manner, the point P in space at which \wp is located. In the case of a Riemann surface, the only difference is that the Riemann surface and the embedding plane have the same dimension. To abstraction from the embedding there corresponds, on the algebraic side, the viewpoint of invariance under arbitrary birational transformations. To enter the realm of topology we must ignore the conformal structure associated with the free-floating Riemann surface. Continuing the comparison, we can say that the conformal structure of the Riemann surface is the equivalent of the metric structure of an ordinary surface, controlled by the first fundamental form, or of the affine and projective structures associated with surfaces in affine and projective differential geometry respectively. In the continuum of real numbers, it is the algebraic operations of $+$ and \cdot that reflect its structural aspect, and in a continuous group the law that associates with an ordered pair of elements their product plays an analogous role. These comments may have increased our appreciation of the relation of the methods. It is a question of rank, of what is viewed as primary. In topology we begin with the notion of continuous connection, and in the course of specialization we add, step by step, relevant structural features. In algebra this order is, in a sense, reversed. Algebra views the operations as the beginning of all mathematical thinking and admits continuity, or some algebraic surrogate of continuity, at the last step of specialization. The two methods follow opposite directions. Little wonder that they don't get on well together. What is most easily accessible to one is often most deeply hidden to the other. In the last few years, in the theory of representation of continuous groups by means of linear substitutions, I have experienced most poignantly how difficult it is to serve these two masters at the same time. Such classical theories as that of algebraic functions can be made to fit both viewpoints. But viewed from these two viewpoints they present completely different sights.

PROBLEMS AND SOLUTIONS

Edited by:

Richard T. Bumby, Fred Kochman and Douglas B. West

Proposed problems should be sent to the MONTHLY PROBLEMS address given on the inside front cover. Please include solutions, relevant references, etc. Three copies are requested.

Solutions of published problems should arrive before October 31, 1995 at the MONTHLY PROBLEMS address given on the inside front cover. Solutions should be typed with double spacing, including the problem number and the solver's name and mailing address. Two copies suffice. A self-addressed postcard or label should be included if an acknowledgement is desired.

An asterisk () after the number of a problem, or part of a problem, indicates that no solution is currently available. Partial solutions will be useful in such cases. Otherwise, the published solution is likely to be based on a solution which is complete and correct. Of course, an elegant partial solution or a method leading to a more general result is always useful and welcome. In addition, references to other appearances of MONTHLY problems or to solutions of these problems in the literature are also solicited.*

PROBLEMS

10452. *Proposed by Seung-Jin Bang, Ajou University, Suwon, Korea.*

Find all values of n , k , a and b (n and k positive integers, $n > k$, a and b nonzero real numbers) for which the polynomial $x^n + ax + b$ is divisible by $x^k + ax + b$ in $\mathbb{R}[x]$.

10453. *Proposed by Murray S. Klamkin, University of Alberta, Edmonton, Alberta, Canada.*

Prove that the following two properties of the altitudes of an n dimensional simplex are equivalent:

- i) the altitudes are concurrent;
- ii) the feet of the altitudes are the orthocenters of their respective faces.

10454. *Proposed by Harry Tamvakis (student), The University of Chicago, Chicago, IL.*

We say that a natural number n is *amenable* if there exist integers a_1, a_2, \dots, a_n such that

$$a_1 + a_2 + \dots + a_n = a_1 a_2 \dots a_n = n.$$

Find all amenable natural numbers.

10455. *Proposed by Zachary Franco, Texas A&M University, Kingsville, TX.*

It is easily seen that a parabola can intersect a circle in at most 4 points.

(a) Show that there is a number R such that a regular polygon (of any number of sides) can intersect a parabola in at most R points.

(b)* Find the smallest R with this property.

10456. *Proposed by Daniel B. Shapiro, Ohio State University, Columbus, OH.*

Denote the group of invertible n by n matrices with entries in the complex numbers by $\text{GL}(n, \mathbb{C})$. Two such matrices f and g will be said to *anticommute* if $fg = -gf$. Also let I denote the identity matrix, which is the unit element of this group.

(a) If $n = 2^m n_0$ with n_0 odd, show that there are k elements of $\text{GL}(n, \mathbb{C})$ that anticommute pairwise if and only if $k \leq 2m + 1$.

(b) If $n = 2^m$ and f_1, \dots, f_{2m} anticommute pairwise, show that the set of products $f_{i_1} f_{i_2} \cdots f_{i_s}$ with $1 \leq i_1 < \dots < i_s \leq 2m$ forms a basis for the 2^{2m} dimensional space of all n by n matrices over \mathbb{C} . Moreover, in this case each f_i^2 is a scalar matrix.

10457. *Proposed by Henry Cohn (student), Massachusetts Institute of Technology, Cambridge, MA.*

Determine the simple continued fraction of $\left(\frac{F_{10n+1}}{F_{10n}} \right)^5$, where F_k denotes the k -th Fibonacci number.

10458. *Proposed by Frank Schmidt, Arlington, VA, and Louis W. Shapiro, Howard University, Washington, DC.*

For a finite group G , let $\text{cd}(G)$ denote the multiset of irreducible character degrees, and let $\text{CD}(G)$ denote the underlying set of distinct character degrees.

(a) Find all n for which the multiset $\{1, 2, \dots, n\}$ appears as $\text{cd}(G)$ for some group G .

(b) Find an upper bound for those n for which the set $\{1, 2, \dots, n\}$ appears as $\text{CD}(G)$ for some group G .

NOTES

(10457) To get started, here are the results for $n = 1$: $F_{10} = 55$ and $F_{11} = 89$, so $F_{11}^5 = 5584059449$ and $F_{10}^5 = 503284375$, and the continued fraction of their ratio is $[11, 10, 1, 1, 2188, 1, 17, 10, 1, 4, 11]$.

SOLUTIONS

An Algebraic Polar Decomposition

6668 [1991,767]. *Proposed by Dragomir Ž. Đoković, University of Waterloo, Ontario, Canada.*

If n is even, show that there exists $A \in \text{GL}_n(\mathbb{C})$ which cannot be represented in the form $A = QS$ with $Q \in \text{SO}_n(\mathbb{C})$ and S symmetric. (For every n it is known that every $A \in \text{GL}_n(\mathbb{C})$ can be represented as $A = QS$ with $Q \in \text{O}_n(\mathbb{C})$ and S symmetric.)

Solution by the proposer. Let $M = \begin{pmatrix} 2 & i \\ i & 0 \end{pmatrix}$. The minimal polynomial of M is $g(\lambda) = (\lambda - 1)^2$, so its Jordan canonical form is $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$. Let T be the n by n matrix formed from m copies (where $n = 2m$) of M along the diagonal and zeros elsewhere. Then, every eigenvalue of T is $+1$ and the Jordan canonical form of T has m 2-by-2 blocks. Now, let $A = DT$, where $D = \text{diag}(-1, 1, 1, \dots, 1)$. Note that $\det(A) = -1$.

Suppose that $A = QS$ with $Q \in O_n(\mathbb{C})$ and S symmetric. Then $S^2 = A^T A = T^2$, so every eigenvalue of S is ± 1 . Using the known Jordan canonical form of T , one sees that T^2 also has Jordan blocks of only even size. This implies that the same is true of S . In particular, every eigenvalue of S has even multiplicity. Thus, $\det(S) = +1$, so that $\det(Q) = -1$. Thus $Q \notin SO_n(\mathbb{C})$, as required.

Editorial comment. The proposer noted that it is classical that for $A \in GL_n(\mathbb{C})$, one has $A = QS$ as above. A proof can be found in F. R. Gantmacher, *The Theory of Matrices*, Chelsea, 1960, chap. XI, §2. He also noted that this theory was extended to singular matrices in I. Kaplansky, "Algebraic polar decomposition", *SIAM J. Matrix Anal. Appl.* 11 (1990), 213–217, where it is proved that a square matrix A over any algebraically closed field of characteristic different from 2 can be written as QS if and only if $(AA^T)^k$ and $(A^T A)^k$ have the same rank for all k .

Robin J. Chapman did not find an element of $GL_n(\mathbb{C})$, but showed that if the singular matrix $B = D(T - I)$ is written in the form QS , then $\det(Q) = -1$. His method used the equation $S^2 = B^T B = (T - I)^2 = 0$ to show that the kernel and image of S and the kernel and image of $T - I$ are all the same subspace of dimension m , and the inner product of any two elements of this subspace is zero. Then the orthogonal matrix $D^{-1}Q$ takes this subspace to itself. This forces $D^{-1}Q \in SO_n(\mathbb{C})$.

No other solutions were received.

A Variant of Prince Rupert's Problem

10251 [1992, 782]. *Proposed by J. G. Mauldon, Amherst College, Amherst, MA.*

Let \mathcal{C} denote the unit cube, and let \mathcal{P} be the set of all pairs $[\mathbf{a}, \mathbf{b}]$ with \mathbf{a} and \mathbf{b} mutually perpendicular line segments contained in \mathcal{C} .

(a) Evaluate $\sup \left\{ \min\{|\mathbf{a}|, |\mathbf{b}|\} : [\mathbf{a}, \mathbf{b}] \in \mathcal{P} \right\}$.

(b) Deduce the area of the largest square, and the volume of the largest regular octahedron, that fit into \mathcal{C} .

Solution by Robin J. Chapman, University of Exeter, Exeter, U. K. The answers are:

(a) $\sup\{\min\{|\mathbf{a}|, |\mathbf{b}|\} : [\mathbf{a}, \mathbf{b}] \in \mathcal{P}\} = 3/2$.

(b) The largest square fitting into \mathcal{C} has area $9/8$, and the largest regular octahedron fitting into \mathcal{C} has volume $9/16$.

Proof. For convenience, we shall instead consider a cube \mathcal{C}' centered at the origin and having vertices $(\pm 1, \pm 1, \pm 1)$. Consider the three line segments \mathbf{a} , \mathbf{b} and \mathbf{c} centered at the origin and having respectively an endpoint at $(1/2, 1, 1)$, $(-1, -1/2, 1)$ and $(-1, 1, -1/2)$. These segments are mutually perpendicular and each has length 3.

To prove (a) we must show that two perpendicular line segments contained in \mathcal{C}' cannot both have lengths exceeding 3. Assuming this for the moment, we prove (b). Any square contained in \mathcal{C}' has two perpendicular diagonals of length d , say, and thus has area $d^2/2$. It follows that the square with diagonals \mathbf{a} and \mathbf{b} has area $9/2$ and is as large as possible among squares in \mathcal{C}' . The largest square in \mathcal{C} thus has area $9/8$. Similarly, any regular octahedron contained in \mathcal{C}' has three mutually perpendicular diagonals of length d , say, and

thus volume $d^3/6$. The octahedron with diagonals **a**, **b** and **c** thus has volume $27/6 = 9/2$, and the largest octahedron in \mathcal{C} has volume $9/16$.

If **x** and **y** are in \mathcal{C}' then the coordinates of **x** – **y** have absolute value at most 2, and so $\pm(\mathbf{x} - \mathbf{y})/2 \in \mathcal{C}'$. It follows that given a line segment **a** contained in \mathcal{C}' there is a parallel line segment **a'** through the origin also contained in \mathcal{C}' . Hence we may restrict attention to line segments through the origin, and as we can extend such segments to the boundary, we may restrict attention to those line segments through the origin whose endpoints lie on the surface of the cube.

Suppose that we fix such a segment **x** of length exceeding 3. By symmetry, we can assume that an endpoint of **x** is $(1, a, b)$ with $0 \leq a \leq b \leq 1$. Since $|\mathbf{x}| > 3$, it follows that $a > 1/2$. If **y** is a segment perpendicular to **x**, we need to consider three possibilities for an endpoint *P* of **y**: $(-1, u, v)$, $(u, -1, v)$ and $(u, v, -1)$, where $|u|, |v| \leq 1$.

First suppose that $P = (-1, u, v)$. Since **y** is perpendicular to **x**, we have $au + bv = 1$. In the (u, v) plane, we are considering points on this line in the square $|u|, |v| \leq 1$, and we want to maximize $u^2 + v^2$ so as to find the maximum possible length for **y**. This maximum occurs where the line meets the boundary of the square, and this happens at the points $u = 1, v = (1 - a)/b$ and $u = (1 - b)/a, v = 1$. It is easy to check that $1 \geq (1 - a)/b \geq (1 - b)/a \geq 0$ when $1/2 < a \leq b \leq 1$, and this gives

$$|\mathbf{y}|^2 = 4(1 + u^2 + v^2) \leq 4\left(2 + \left(\frac{1-a}{b}\right)^2\right).$$

If $|\mathbf{y}| > 3$, this yields $(1 - a)/b > 1/2$ and thus $a < 1 - b/2$. We also have $a^2 + b^2 > 5/4$ since $|\mathbf{x}| > 3$, and this yields

$$\left(1 - \frac{b}{2}\right)^2 + b^2 > a^2 + b^2 > \frac{5}{4}.$$

There is no solution to this inequality in the range $0 \leq b \leq 1$.

Next, suppose $P = (u, -1, v)$. Since **y** is perpendicular to **x**, $u + bv = a$. In the (u, v) -plane, this meets the boundary of the square $|u|, |v| \leq 1$ at the points $u = 1, v = (a - 1)/b$ and $u = a = b, v = 1$. In the range $0 < a \leq b \leq 1$, we always have $(1 - a)/b \geq b - a \geq 0$, and thus

$$|\mathbf{y}|^2 = 4(1 + u^2 + v^2) \leq 4\left(2 + \left(\frac{1-a}{b}\right)^2\right).$$

As in the previous case, this leads to a contradiction if $|\mathbf{y}| > 3$.

Finally, suppose $P = (u, v, -1)$. Then $u + av = b$ and the points on the boundary of the square are $u = 1, v = (b - 1)/a$ and $u = b - a, v = 1$. This time, both points must be considered. As before, $|\mathbf{y}| > 3$ requires that either $b - a > 1/2$ or $(1 - a)/b > 1/2$, while $a^2 + b^2 > 5/4$. If $b - a > 1/2$, we have

$$\left(b - \frac{1}{2}\right)^2 + b^2 > \frac{5}{4}$$

and if $(1 - b)/a > 1/2$, we have

$$4(1 - b)^2 + b^2 > \frac{5}{4}.$$

Neither of these quadratic inequalities has solutions for $1/2 \leq b \leq 1$.

Editorial comment. The problem of finding the largest square inscribed in a unit cube is quite old. In H. Croft, K. Falconer & R. Guy, *Unsolved Problems in Geometry*, Springer, 1991, it appears as Problem B4 with some unsolved generalizations. The original question is attributed to Prince Rupert (1619–1682), with a solution by Pieter Nieuwland in the

Eighteenth Century. The inequality of part (a) allows a proof not overly dependent on geometric intuition.

Raphael M. Robinson generalized part (a) to two perpendicular segment in a unit n -cube. The desired supremum d_n is

$$d_n = \begin{cases} \sqrt{n} & \text{if } n \text{ is even} \\ \sqrt{n - 3/4} & \text{if } n \text{ is odd.} \end{cases}$$

Solved also by I. Kastanas, O. P. Lossers (The Netherlands), R. M. Robinson, Anchorage Math Solutions Group, and the proposer.

A Fibonacci Series

10262 [1992, 873]. *Proposed by Dean Clark, University of Rhode Island, Kingston, RI.*

Evaluate

$$\sum_{n=1}^{\infty} \frac{(-1)^n}{F_n F_{n+2}}$$

where $\langle F_n \rangle$ denotes the sequence of Fibonacci numbers.

Composite solution and generalization by Anatoly S. Izotov, Mining Institute, Novosibirsk, Russia, the University of Wyoming Problem Circle, and the editors. The answer is $2 - \sqrt{5}$. More generally, suppose $\langle p \rangle$ is a sequence satisfying $p_n = ap_{n-1} + p_{n-2}$ for $n \geq 2$, where $a > 0$. If $k \geq 1$ and $p_k p_1 - p_{k+1} p_0 \neq 0$, we prove that $\lim_{n \rightarrow \infty} \frac{p_{n-1}}{p_n} = \frac{\sqrt{a^2+4}-a}{2}$ and that

$$\sum_{n=1}^{\infty} \frac{(-1)^n}{p_n p_{n+k}} = \frac{1}{p_k p_1 - p_{k+1} p_0} \left(\sum_{n=1}^k \frac{p_{n-1}}{p_n} - k \left(\frac{\sqrt{a^2+4}-a}{2} \right) \right).$$

From the characteristic equation for the recurrence, we have $p_n = A\alpha^n + B\beta^n$, where A, B are constants and $\alpha = \frac{a+\sqrt{a^2+4}}{2}$, $\beta = \frac{a-\sqrt{a^2+4}}{2}$. Since $p_k p_1 - p_{k+1} p_0 \neq 0$, neither A nor B is zero. Because $|\beta/\alpha| < 1$, we have $\lim_{n \rightarrow \infty} \frac{p_{n-1}}{p_n} = \frac{1}{\alpha} = \frac{\sqrt{a^2+4}-a}{2}$.

For fixed $k \geq 0$, it follows easily by induction on n that

$$p_{n-1} p_{n+k} - p_{n+k-1} p_n = (-1)^n (p_k p_1 - p_{k+1} p_0)$$

for $n \geq 1$. Thus,

$$\begin{aligned} \sum_{n=1}^N \frac{(-1)^n}{p_n p_{n+k}} &= \sum_{n=1}^N \frac{1}{p_k p_1 - p_{k+1} p_0} \frac{p_{n-1} p_{n+k} - p_{n+k-1} p_n}{p_n p_{n+k}} \\ &= \frac{1}{p_k p_1 - p_{k+1} p_0} \sum_{n=1}^N \left(\frac{p_{n-1}}{p_n} - \frac{p_{n+k-1}}{p_{n+k}} \right) \\ &= \frac{1}{p_k p_1 - p_{k+1} p_0} \left(\sum_{n=1}^k \frac{p_{n-1}}{p_n} - \sum_{n=N+1}^{N+k} \frac{p_{n-1}}{p_n} \right). \end{aligned}$$

Taking the limit as n approaches infinity completes the proof.

When $k = 2$ and $\langle p \rangle$ is the Fibonacci sequence with $F_1 = F_2 = 1$, the value of the series reduces to $2 - \sqrt{5}$.

Editorial comment. The argument here follows the same steps that most solvers used directly to sum the series posed in the problem statement. Seven solvers mentioned the generalization for arbitrary k when $\langle p \rangle$ is the Fibonacci series, found in Brother Alfred

Brousseau, "Summation of infinite Fibonacci series", *Fibonacci Quarterly*, 7(1969), 143-168. Seung-Jin Bang cited Br. J. M. Mahon and Alwyn F. Horadam, "Infinite series summation involving reciprocals of Pell polynomials", *Fibonacci Numbers and Their Applications*, (A.N. Philippou, G.E. Bergum, and A.F. Horadam, eds.), D.Reidel, 1986, p.168, which studied series determined by other recurrences.

Solved by 80 readers and the proposer. Two incorrect solutions were received.

Largest Product of Distances to Vertices

10282 [1993, 184]. *Proposed by Paul Erdős, Hungarian Academy of Sciences, Budapest, Hungary.*

Let A, B, C be the vertices of a triangle inscribed in a unit circle, and let P be a point in the interior of the triangle ABC . Show that

$$|PA| \cdot |PB| \cdot |PC| < \frac{32}{27}.$$

Solution I by B. J. Venkatachala, Indian Institute of Science, Bangalore, India. Suppose P is a point on the boundary of the triangle ABC , say on the side BC . Let x be the distance of P from the center of the unit circle that circumscribes the triangle ABC . It is well known that, for all chords XY through the point P , the product $|PX| \cdot |PY|$ is constant. Therefore, $|PB| \cdot |PC| = (1-x)(1+x)$ and $|PA| \leq 1+x$. Thus

$$|PA| \cdot |PB| \cdot |PC| \leq (1-x)(1+x)^2.$$

The expression on the right assumes its maximum value $32/27$ in the interval $[0, 1]$ at $x = 1/3$. Finally, by the maximum modulus principle, $|PA| \cdot |PB| \cdot |PC| < 32/27$ for any point P in the interior of the triangle ABC .

Solution II by Harry D. Ruderman, Bronx, NY. Let T be the product of the distances from P to the vertices A, B, C . Slide A along the circumference to increase $|PA|$, if possible. Either A can reach a point where $|PA|$ attains a local maximum, in which case P, O, A are collinear, or it reaches a point where further movement will lead to a triangle not containing P in its interior, in which case P will be on one of the sides AB or AC . Repeating this operation for the other vertices will lead to a triangle in which P lies on a side of ABC , say BC and P, O, A are collinear. If O is not between A and P , replace A by the opposite end of the diameter through P , increasing T . Now, rotate BC about P until it is perpendicular to \overline{AOP} . This does not change T . With $x = |OP|$, the Pythagorean theorem gives $|PB| = |PC| = \sqrt{1-x^2}$, and $T = (1+x)(1-x^2)$ which is maximal when $x = 1/3$ and $T = (4/3)(8/9) = 32/27$.

Editorial comment. The solutions above are the most efficient of the two principle types received. Restriction to the case in which P is on the boundary avoids the need to express the product of the distances from a general point. Restriction to symmetric configurations, as in solution II, was a common approach. Note that configurations attaining the maximal value are easily characterized.

A generalization to an n -gon inscribed in a circle was obtained by: Roy Barbara, Kevin Brown, Hans Georg Killinbergtrø & Ivar Skau, Murray S. Klamkin, Neela Lakshmanan, Arthur J. Rosenthal & Radha G. Nath, Ossama A. Saleh & Terry J. Walters, and Ajaj Tarabay. The bound in this case is the maximum value of $(1-x)(1+x)^{n-1}$ which is $(2/n)^n (n-1)^{n-1}$. A generalization to higher dimensions was given by Faruk F. Abi-Khuzam & Ajaj Tarabay, and a generalization to weighted products for polygons was given by Alladi Ramakrishnan.

This problem has appeared as problem 1895 in *Crux Mathematicorum* (December 1993), and it has been reported that the problem also appears in Bull. Math. (Wuhan) 1990, no.3 p. 17 with solution in 1991, no.10, p.42.

The solution by Harry D. Ruderman was received on February 27, 1993. This was the last solution that we received from him, and he died a little over a year later. He was an active contributor to the Problem Section since joining the Association as a student in 1931, beginning with a published solution to problem 299 [1914, 297; 1931,172; 1931, 462]. Other notable early contributions were solutions (not published in full, but summarized) to 3746 [1935, 454; 1937, 400] and 3848 [1937, 667; 1940, 575], which were geometry problems proposed by Paul Erdős.

Solved by 61 readers (including those cited) and the proposer.

Collaborating editors: David F. Appleyard, Paul T. Bateman, Bruce C. Berndt, Duane M. Broline, Barry W. Brunson, Frank S. Cater, Gulbank D. Chakerian, Underwood Dudley, Gerald A. Edgar, Michael A. Filaseta, Ira M. Gessel, Richard A. Gibbs, Jerrold R. Griggs, Douglas A. Hensley, John R. Isbell, Mourad E. H. Ismail, Murray Klamkin, Daniel J. Kleitman, Frederick W. Luttman, Frank B. Miles, Richard Pfeifer, Stephen L. Portnoy, J. O. Shallit, John Henry Steelman, Kenneth B. Stolarsky, David E. Tepper, Douglas B. Tyler, Daniel Ullman, and William E. Watkins.

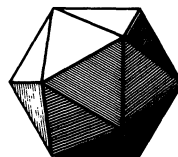
Raphael M. Robinson

In the February 1995 issue of the Monthly, the article by Deborah Haimo, *Experimentation and Conjecture Are Not Enough*, contained a reference to Abraham Robinson and the Banach Tarski Paradox. That paragraph should have read:

Banach and Tarski succeeded in showing the fascinating and nonintuitive result that, say, the earth can be decomposed into a finite number of pieces which can be re-assembled to form a marble, or even to form two earths each of the same size as the original. John von Neumann added to these amazing facts the observation that only nine pieces are needed for the decomposition of one sphere into two, all with the same radii. *Raphael M. Robinson* went further yet in 1947 showing that five pieces will suffice!

With sadness, we also report the death of Raphael M. Robinson on January 27, 1995. He was a longtime member of the MAA, and a valued friend of the Monthly. We will miss his friendly letters and advice.

The American Mathematical Monthly



Volume 102, Number 6/JUNE-JULY 1995



Alan Turing
(see p. 483)

AN OFFICIAL PUBLICATION OF THE MATHEMATICAL ASSOCIATION OF AMERICA

NOTICE TO AUTHORS

The *Monthly* publishes articles, notes, and other features about mathematics and the profession. The readership of the *Monthly* is intended to include everybody who is mathematically inclined, including of course professional mathematicians and students of mathematics at all collegiate levels. While no single article or feature is likely to appeal to everyone, material should interest and be accessible to a large number of readers. This is the most important criterion for acceptance.

Articles may be expositions of old results or presentations of new ones. They may concern all of mathematics or one small area, a broad development or a single application, historical reminiscences or one important event. While some articles may contain the author's new research, the novelty of material and generality of the results is far less important than the clarity of exposition and general interest. Discussing one illuminating case of a well known result is far better than providing all the details of an obscure but new proposition. Articles in the *Monthly* are supposed to inform and to entertain; they are meant to be read rather than archived.

Notes are short and possibly informal articles. A note may concern a clever new proof of an old theorem, a novel way to present tired material, or a lively discussion of a philosophical (but still mathematical) issue. Also, any topic is suitable, so long as it is related to mathematics. Because a note is short, the first few sentences are the most important part: They should explain the purpose and invite the reader in. Photographs or diagrams often will attract the reader's attention.

All articles and notes should be sent to the editor:

JOHN EWING
Department of Mathematics
Indiana University
Bloomington, IN 47405

Please send 3 copies, typewritten on only one side of the paper. Illustrations should be carefully drawn on separate sheets of paper in black ink; the original should be without lettering and two copies should have appropriate captions and lettering indicated.

Proposed problems or solutions should be sent to:

RICHARD BUMBY,
P.O. Box 10971
New Brunswick, NJ 08906-0971.

Please send 2 copies of all material, typewritten if possible.

Letters to the Editor, both for publication and for private reading, should be sent to the Editor at the address given above. Comments, including criticisms, are welcome, as are all suggestions for making the *Monthly* a lively, entertaining, and informative journal.

EDITOR:

JOHN H. EWING

ASSOCIATE EDITORS:

PETER BORWEIN	FRED KOCHMAN
RICHARD BUMBY	CATHERINE MCGEOCH
DENNIS DETURCK	RICHARD NOWAKOWSKI
UNDERWOOD DUDLEY	ARNOLD OSTEBEE
JOHN DUNCAN	LEE RUBEL
JOAN FERRINI-MUNDY	ABE SHENITZER
JOSEPH GALLIAN	LYNN STEEN
STEVEN GALOVICH	STAN WAGON
RICHARD GUY	DOUGLAS WEST
DARRELL HAILE	HERBERT WILF
PAUL HALMOS	SANDY ZABELL
JOAN HUTCHINSON	PAUL ZORN

EDITORIAL ASSISTANT:

MISTY CUMMINGS

STAFF ARTIST:

MIKE CAGLE

QUOTE MASTER:

MARK WOODARD

Reprint permission:

MARCIA P. SWARD, Executive Director

Advertising Correspondence:

Ms. ELAINE PEDREIRA, Advertising Manager

Subscription correspondence, change of address, and other inquiries:

Membership / Subscriptions Department

All at the address:

The Mathematical Association of America
1529 Eighteenth Street, N.W.
Washington, DC 20036.

Microfilm Editions: University Microfilms International, Serial Bid coordinator, 300 North Zeeb Road, Ann Arbor, MI 48106.

The AMERICAN MATHEMATICAL MONTHLY (ISSN 0002-9890) is published monthly except bimonthly June-July and August-September by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, DC 20036 and Montpelier, VT. Copyrighted by the Mathematical Association of America (Incorporated), 1995, including rights to this journal issue as a whole and, except where otherwise noted, rights to each individual contribution. General permission is granted to Institutional Members of the MAA for noncommercial reproduction in limited quantities of individual articles (in whole or in part) provided a complete reference is made to the source. Second class postage paid at Washington, DC, and additional mailing offices. **Postmaster:** Send address changes to the American Mathematical Monthly, Membership / Subscription Department, MAA, 1529 Eighteenth Street, N.W., Washington, DC, 20036-1385.

**The American
Mathematical Monthly**

Volume 102 Number 6 / JUNE–JULY 1995
(ISSN 0002-9890)



Contents

ARTICLES

- Alan Turing and the Central Limit Theorem / S. L. ZABELL 483
Niels Hendrik Abel and Equations of the Fifth Degree /
MICHAEL I. ROSEN 495
The Great Marble Race: An Assignment Gone Wrong / BENNY EVANS and
JERRY JOHNSON 506
Geometry and the Foucault Pendulum / JOHN OPREA 515
Areas of Polygons Inscribed in a Circle / DAVID P. ROBBINS 523
Curves of Constant Precession / PAUL D. SCOFIELD 531
-

FEATURES

COMMENTS 482

NOTES

- Matrix Expansion by Orthogonal Kronecker Products /
JEFFERY C. ALLEN 538
Injective Polynomial Maps are Automorphisms /
WALTER RUDIN 540
An Elementary Proof of the Simplicity of the Mathieu Groups M_{11}
and M_{23} / ROBIN J. CHAPMAN 544

UNSOLVED PROBLEMS

- Wanted: A Bad Matrix / GARY H. MEISTERS 546

THE AUTHORS 551

PROBLEMS AND SOLUTIONS 553

REVIEWS

- The Words of Mathematics: An Etymological Dictionary of Mathematical
Terms Used in English.* By Steven Schwartzman /
HENRY J. RICARDO 563

TELEGRAPHIC REVIEWS 566

Alan Turing and the Central Limit Theorem

S. L. Zabell

Although the English mathematician Alan Mathison Turing (1912–1954) is remembered today primarily for his work in mathematical logic (Turing machines and the “Entscheidungsproblem”), machine computation, and artificial intelligence (the “Turing test”), his name is not usually thought of in connection with either probability or statistics. One of the basic tools in both of these subjects is the use of the normal or Gaussian distribution as an approximation, one basic result being the Lindeberg-Feller central limit theorem taught in first-year graduate courses in mathematical probability. No one associates Turing with the central limit theorem, but in 1934 Turing, while still an undergraduate, rediscovered a version of Lindeberg’s 1922 theorem and much of the Feller-Lévy converse to it (then unpublished). This paper discusses Turing’s connection with the central limit theorem and its surprising aftermath: his use of statistical methods during World War II to break key German military codes.

1. INTRODUCTION. Turing went up to Cambridge as an undergraduate in the Fall Term of 1931, having gained a scholarship to King’s College. (Ironically, King’s was his second choice; he had failed to gain a scholarship to Trinity.) Two years later, during the course of his studies, Turing attended a series of lectures on the Methodology of Science, given in the autumn of 1933 by the distinguished astrophysicist Sir Arthur Stanley Eddington. One topic Eddington discussed was the tendency of experimental measurements subject to errors of observation to often have an approximately normal or Gaussian distribution. But Eddington’s heuristic sketch left Turing dissatisfied; and Turing set out to derive a rigorous mathematical proof of what is today termed the central limit theorem for independent (but not necessarily identically distributed) random variables.

Turing succeeded in his objective within the short span of several months (no later than the end of February 1934). Only then did he find out that the problem had already been solved, twelve years earlier, in 1922, by the Finnish mathematician Jarl Waldemar Lindeberg (1876–1932). Despite this, Turing was encouraged to submit his work, suitably amended, as a Fellowship Dissertation. (Turing was still an undergraduate at the time; students seeking to become a Fellow at a Cambridge college had to submit evidence of original work, but did not need to have a Ph.D. or its equivalent.) This revision, entitled “On the Gaussian Error Function,” was completed and submitted in November, 1934. On the strength of this paper Turing was elected a Fellow of King’s four months later (March 16, 1935) at the age of 22; his nomination supported by the group theorist Philip Hall and the economists John Maynard Keynes and Alfred Cecil Pigou. Later that year the paper was awarded the prestigious Smith’s prize by the University (see Hodges, 1983).

Turing never published his paper. Its major result had been anticipated, although, as will be seen, it contains other results that were both interesting and novel at the time. But in the interim Turing's mathematical interests had taken a very different turn. During the spring of 1935, awaiting the outcome of his application for a Fellowship at King's, Turing attended a course of lectures by the topologist M. H. A. Newman on the Foundations of Mathematics. During the International Congress of Mathematicians in 1928, David Hilbert had posed three questions: is mathematics *complete* (that is, can every statement in the language of number theory be either proved or disproved?), is it *consistent*, and is it *decidable*? (This last is the *Entscheidungsproblem*, or the "decision problem"; does there exist an *algorithm* for deciding whether or not a specific mathematical assertion does or does not have a proof.) Kurt Gödel had shown in 1931 that the answer to the first question is *no* (the so-called "first incompleteness theorem"); and that if number theory is consistent, then a proof of this fact does not exist using the methods of the first-order predicate calculus (the "second incompleteness theorem"). Newman had proved the Gödel theorems in his course, but he pointed out that the third of Hilbert's questions, the *Entscheidungsproblem*, remained open.

This challenge attracted Turing, and in short order he had arrived at a solution (in the negative), using the novel device of *Turing machines*. The drafting of the resulting paper (Turing, 1937), dominated Turing's life for a year from the Spring of 1935 (Hodges, 1983, p. 109); and thus Turing turned from mathematical probability, never to return.

A copy of Turing's Fellowship Dissertation survives, however, in the archives of the King's College Library; and its existence raises an obvious question. Just how far did a mathematician of the calibre of Turing get in this attack on the central limit theorem, one year before he began his pioneering research into the foundations of mathematical logic? The answer to that question is the focus of this paper.

2. THE CENTRAL LIMIT THEOREM. The earliest version of the central limit theorem (CLT) is due to Abraham de Moivre (1667–1754). If X_1, X_2, X_3, \dots is an infinite sequence of 1's and 0's recording whether a success ($X_n = 1$) or failure ($X_n = 0$) has occurred at each stage in a sequence of repeated trials, then the sum $S_n = X_1 + X_2 + \dots + X_n$ gives the total number of successes after n trials. If the trials are independent, and the probability of a success at each trial is the same, say $P[X_n = 1] = p$, $P[X_n = 0] = 1 - p$, then the probability of seeing exactly k successes in n trials has a binomial distribution:

$$P[S_n = k] = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}.$$

If n is large (for example, 10,000), then as de Moivre noted, the direct computation of binomial probabilities "is not possible without labor nearly immense, not to say impossible"; and for this reason he turned to approximate methods (see Diaconis and Zabell, 1991): using Stirling's approximation (including correction terms) to estimate the individual terms in the binomial distribution and then summing, de Moivre discovered the remarkable fact that

$$\lim_{n \rightarrow \infty} P \left[a \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq b \right] = \frac{1}{\sqrt{2\pi}} \int_a^b \exp \left[-\frac{1}{2} x^2 \right] dx,$$

or $\Phi(b) - \Phi(a)$, where $\Phi(x)$ is the cumulative distribution function of the standard normal (or Gaussian) distribution:

$$\Phi(x) =: \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp\left[-\frac{1}{2}t^2\right] dt.$$

During the 19th and 20th centuries this result was extended far beyond the simple coin-tossing setup considered by de Moivre, important contributions being made by Laplace, Poisson, Chebyshev, Markov, Liapunov, von Mises, Lindeberg, Lévy, Bernstein, and Feller; see Adams (1974), Maistrov (1974), Le Cam (1986), and Stigler (1986) for further historical information. Such investigations revealed that if X_1, X_2, X_3, \dots is *any* sequence of independent random variables having the same distribution, then the sum S_n satisfies the CLT provided suitable centering and scaling constants are used: the centering constant np in the binomial case is replaced by the sum of the *expectations* $E[X_n]$; the scaling constant $\sqrt{np(1-p)}$ is replaced by the square root of the sum of the *variances* $\text{Var}[X_n]$ (provided these are finite).

Indeed, it is not even necessary for the random variables X_n contributing to the sum S_n to have the same distribution, provided that no one term dominates the sum. Of course this has to be made precise. The best result is due to Lindeberg. Suppose $E[X_n] = 0$, $0 < \text{Var}[X_n] < \infty$, $s_n^2 =: \text{Var}[S_n]$, and

$$\Lambda_n(\varepsilon) =: \sum_{k=1}^n E\left[\left(\frac{X_k}{s_n}\right)^2; \frac{|X_k|}{s_n} \geq \varepsilon\right].$$

(The notation $E[X; Y \geq \varepsilon]$ means the expectation of X is restricted to outcomes ω such that $Y(\omega) \geq \varepsilon$.) The *Lindeberg condition* is the requirement that

$$\Lambda_n(\varepsilon) \rightarrow 0, \quad \forall \varepsilon > 0; \quad (2.1)$$

and the *Lindeberg central limit theorem* (Lindeberg, 1922) states that if the sequence of random variables X_1, X_2, \dots satisfies the Lindeberg condition (2.1), then for all $a < b$,

$$\lim_{n \rightarrow \infty} P\left[a < \frac{S_n}{s_n} < b\right] = \Phi(b) - \Phi(a). \quad (2.2)$$

Despite its technical appearance, the Lindeberg condition turns out to be a natural sufficient condition for the CLT. There are two reasons for this. First, the Lindeberg condition has a simple consequence: if $\sigma_k^2 =: \text{Var}[X_k]$, then

$$\rho_n^2 =: \max_{k \leq n} \left(\frac{\sigma_k^2}{s_n^2}\right) \rightarrow 0. \quad (2.3)$$

Thus, if the sequence X_1, X_2, X_3, \dots satisfies the Lindeberg condition, the variance of an individual term X_k in the sum S_n is asymptotically negligible. Second, for such sequences the Lindeberg condition is *necessary* as well as sufficient for the CLT to hold, a beautiful fact discovered (independently) by William Feller and Paul Lévy in 1935. In short: (2.1) \leftrightarrow (2.2) + (2.3).

If, in contrast, the Feller-Lévy condition (2.3) fails, then it turns out that convergence to the normal distribution can occur in a fashion markedly different from that of the CLT. If (2.3) does *not* hold, then there exists a number $\rho > 0$, and

two sequences of positive integers $\{m_k\}$ and $\{n_k\}$, $\{n_k\}$ is strictly increasing such that

$$1 \leq m_k \leq n_k \quad \text{for all } k \text{ and } \text{Var} \left[\frac{X_{m_k}}{s_{n_k}} \right] = \frac{\sigma_{m_k}^2}{s_{n_k}^2} \rightarrow \rho^2 > 0. \quad (2.4)$$

Feller (1937) showed that if normal convergence occurs (that is, condition (2.2) holds), but condition (2.4) also obtains, then

$$\frac{1}{\rho} \frac{X_{m_k}}{s_{n_k}} \Rightarrow N(0, 1).$$

That is, there exists a subsequence X_{m_k} whose contributions to the sums S_n are nonnegligible (relative to s_n) and which, properly scaled, converges to the standard normal distribution. (The symbol “ \Rightarrow ” denotes convergence in distribution; $N(\mu, \sigma^2)$ the normal distribution having expectation μ , variance σ^2 .)

Note. For the purposes of brevity, this summary of the contributions of Feller and Lévy simplifies a much more complex story; see Le Cam (1986) for a more detailed account. (Or better, consult the original papers themselves!)

3. TURING’S FELLOWSHIP DISSERTATION. Turing’s fellowship dissertation was written twelve years after Lindeberg’s work had appeared, and shortly before the work of Feller and Lévy. There are several aspects of the paper that demonstrate Turing’s insight into the basic problems surrounding the CLT. One of these is his decision, contrary to a then common textbook approach (see, e.g., Burnside, 1928, pp. 87–90), but crucial if the best result is to be obtained (and the approach also adopted by Lindeberg), to work at the level of distribution functions (i.e., the function $F_X(t) =: P[X \leq t]$) rather than densities (the derivatives of the distribution functions). In Appendix B Turing notes:

I have attempted to obtain some results [using densities]...but without success. The reason is clear. In order that the shape frequency functions $u_n(x)$ of $f_n(x)$ should tend to the shape frequency function $\phi(x)$ of the Gaussian error much heavier restrictions on the functions $g_n(x)$ are required than is needed if we only require that $U_n \rightarrow \Phi$. It became clear to me...that it would be better to work in terms of distribution function throughout.

This was an important insight. Although versions of the central limit theorem do exist for densities, these ordinarily require stronger assumptions than just the Lindeberg condition (2.1); see, e.g., Feller (1971), pp. 516–517, Petrov (1975), Chapter 7.

Let us now turn to the body of Turing’s paper, and consider it, section by section.

3.1. Basic Structure of the Paper. The first seven sections of the paper (pp. 1–6) summarize notation and the basic properties of distribution functions. Section 1 summarizes the problem; Section 2 defines the distribution function F (abbreviated DF) of an “error” ε ; Section 3 summarizes the basic properties of the expectation and mean square deviation (MSD) of a sum of independent errors; rigorous proofs in terms of the distribution function are given in an appendix at the end of the paper (Appendix C). Section 4 discusses the distribution function of a sum of independent errors, the *sum distribution function* (SDF), in terms of the distribution functions of each term in the sum, and derives the formula for $F \oplus G$,

the convolution of two distribution functions. Section 5 then introduces the concept of the *shape function* (SF); the standardization of a distribution function F to have zero expectation and unit MSD; thus, if F has expectation μ and MSD σ^2 ($\sigma > 0$), then the shape function of F is $U(x) =: F(\sigma(x - \mu))$. (Turing uses the symbols “ a ” and “ k^2 ” to denote μ and σ^2 ; several other minor changes in notation of this sort are made below.)

In Section 6 Turing then states the basic problem to be considered: given a sequence of errors ε_k , having distribution functions G_k , shape functions V_k , means μ_k , mean square deviations σ_k^2 , sum distribution functions F_n , and shape functions U_n for each F_n , under what conditions do the shape functions $U_n(x)$ converge uniformly to $\Phi(x)$, the “SF of the Gaussian Error”? Turing then assumes for simplicity that $\mu_k = 0$ and $\sigma_k^2 < \infty$. In Section 7 (“Fundamental Property of the Gaussian Error”), he notes the only properties of Φ that are used in deriving sufficient conditions for normal convergence are that it is an SF, and the “self-reproductive property” of Φ : that is, if $X_1 \sim N(0, \sigma_1^2)$ and $X_2 \sim N(0, \sigma_2^2)$ are independent, then $X_1 + X_2 \sim N(0, \sigma_1^2 + \sigma_2^2)$. (The notation “ $X \sim N(\mu, \sigma^2)$ ” means that the random variable X has the distribution $N(\mu, \sigma^2)$.)

3.2. The Quasi-Necessary Conditions. It is at this point that Turing comes to the heart of the matter. In Section 8 (“The Quasi-Necessary Conditions”) Turing notes

The conditions we shall impose fall into two groups. Those of one group (the quasi-necessary conditions) involve the MSDs only. They are not actually necessary, but if they are not fulfilled U_n can only tend to Φ by a kind of accident.

The two conditions that Turing refers to as the “quasi-necessary” conditions are:

$$\sum_{k=1}^{\infty} \sigma_k^2 = \infty \quad \text{and} \quad \frac{\sigma_n^2}{s_n^2} \rightarrow 0. \quad (3.1)$$

It is easy to see that Turing’s condition (3.1) is equivalent to condition (2.3). (That (2.3) \Rightarrow (3.1) is immediate. To see (3.1) \Rightarrow (2.3): given $\varepsilon > 0$, choose $M \geq 1$ so that $\sigma_n^2/s_n^2 < \varepsilon$ for $n \geq M$, and $N \geq M$ so that $\sigma_k^2/s_N^2 < \varepsilon$ for $1 \leq k \leq M$; if $n \geq N$, then $\sigma_k^2/s_n^2 < \varepsilon$ for $1 \leq k \leq n$.)

In his Theorems 4 and 5, Turing explores the consequences of the failure of either part of condition (3.1). Turing’s proof of Theorem 4 requires his

Theorem 3. *If X and Y are independent, and both X and $X + Y$ are Gaussian, then Y is Gaussian.*

This is a special case of a celebrated theorem proven shortly thereafter by Harald Cramér (1936); if X and Y are independent, and $X + Y$ is Gaussian, then both X and Y must be Gaussian. Lévy had earlier conjectured Cramér’s theorem to be true (in 1928 and again in 1935), but had been unable to prove it. Cramér’s proof of this result in 1936 in turn enabled Lévy to arrive at necessary and sufficient conditions for the CLT of a very general type (using centering and scaling constants other than the mean and standard deviation), and this in turn led Lévy to write his famous monograph, *Théorie de l’Addition des Variables Aléatoires* (Lévy, 1937); see Le Cam (1986, pp. 80–81, 90).

Cramér’s theorem is a hard fact; his original proof appealed to Hadamard’s theorem in the theory of entire functions. The special case of the theorem needed

by Turing is much simpler; it is an immediate consequence of the characterization theorem for characteristic functions. To see this, let $\phi_X(t) =: E[\exp(itX)]$ denote the characteristic function of a random variable X ; and suppose that X and Y are independent, $X \sim N(0, \sigma^2)$, and $X + Y \sim N(0, \sigma^2 + \tau^2)$. Then

$$\exp\left(-\frac{\sigma^2 + \tau^2}{2}t^2\right) = \phi_{X+Y}(t) = \phi_X(t)\phi_Y(t) = \exp\left(-\frac{\sigma^2}{2}t^2\right)\phi_Y(t),$$

hence $\phi_Y(t) = \exp(-(\tau^2/2)t^2)$; thus $Y \sim N(0, \tau^2)$ because the characteristic function of a random variable uniquely determines the distribution of that variable. Turing's proof, which uses distribution functions, is not much longer.

It is an immediate consequence of Cramér's theorem that if $S_n/s_n \Rightarrow N(0, 1)$, but $\lim_{n \rightarrow \infty} s_n^2 < \infty$, then all the summands X_j must in fact be Gaussian. But Turing did not have this fact at his disposal, only his much weaker Theorem 3. His **Theorem 4** (phrased in the language of random variables) thus makes the much more limited claim that if (a) $\Sigma \sigma_n^2 < \infty$, (b) S_n converges to a Gaussian distribution, and (c) X_0 is a random variable at once independent of the original sequence X_1, X_2, \dots and having a distribution other than Gaussian, then the sequence $S_n^* = X_0 + S_n$ *cannot* converge to the Gaussian distribution. In other words: if $\Sigma \sigma_n^2 < \infty$, then "the convergence... to the Gaussian is so delicate that a single extra term in the sequence... upsets it" (p. 17).

Turing's **Theorem 5** in turn explores the consequences of the failure of (3.1) in the case that $\Sigma \sigma_n^2 = \infty$, but $\rho_n^2 =: \sigma_n^2/s_n^2$ does not tend to zero as $n \rightarrow \infty$. The statement of the theorem is somewhat technical in nature, but Turing's later summary of it captures the essential phenomenon involved:

If F_n [the distribution function of S_n] tends to Gaussian and σ_n^2/s_n^2 does not tend to zero [but $\Sigma \sigma_n^2 = \infty$] we can find a subsequence of G_n [the distribution function of X_n] tending to Gaussian.

Thus Turing had by some two years anticipated Feller's discovery of the subsequence phenomenon. (In Turing's typescript, symbols such as " F_n " are entered by hand; in the above quotation the space for " F_n " has by accident been left blank, but the paragraph immediately preceding this one in the typescript makes it clear that " F_n " is intended.)

3.3 The Sufficient Conditions. Turing states in his preface that he had been "informed that an almost identical proof had been given by Lindeberg." This comment refers to the *method* of proof Turing uses, not the *result* obtained. Turing's method is to smooth the distribution functions $F_n(x)$ of the sum by forming the convolution $F_n * \Phi(x/\rho)$, expand the result in a Taylor series to third order, and then let the variance ρ^2 of the convolution term tend to zero. This is similar to the method employed by Lindeberg. (There is an important difference, however: Turing does not use Lindeberg's "swapping" argument. For an attractive modern presentation of the Lindeberg method, see Breiman, 1968, pp. 167–170; for discussion of the method, Pollard's comments in Le Cam, 1986, pp. 94–95.)

Turing does *not*, however, succeed in arriving at the Lindeberg condition (2.1) as a sufficient condition for convergence to the normal distribution; the most general sufficient condition he gives (on p. 27) is complex in appearance (although it necessarily implies the Lindeberg condition). Turing concedes that his "form of the sufficiency conditions is too clumsy for direct application," but notes that it can be used to "derive various criteria from it, of different degrees of directness and of comprehensiveness" (p. 28). One of these holds if the summands X_k all have the

same *shape* (that is, the shape functions $V_k(x) =: P[X_k/\sigma_k \leq x]$ coincide); and thus includes the special case of identically distributed summands having a second moment. (This was no small feat, since even this special case of the more general Lindeberg result had eluded proof until the publication of Lindeberg's paper.)

One formulation of this criterion, equivalent to the one actually stated by Turing, is: there exists a function $J: \mathbf{R}^+ \rightarrow \mathbf{R}^+$ such that $\lim_{t \rightarrow \infty} J(t) = 0$, and

$$E \left[\left(\frac{X_k}{\sigma_k} - t \right)^2; \left| \frac{X_k}{\sigma_k} \right| \geq t \right] \leq J(t) \quad \text{for all } k \geq 1, t \geq 0. \quad (3.2)$$

In turn one simple sufficient condition for this given by Turing (pp. 30–31) is that there exists a function ϕ such that $\phi(x) > 0$ for all x , $\lim_{x \rightarrow \pm\infty} \phi(x) = \infty$, and

$$\sup_k E \left[\left(\frac{X_k}{\sigma_k} \right)^2 \phi \left(\frac{X_k}{\sigma_k} \right) \right] < \infty. \quad (3.3)$$

(Note that unfortunately one important special case not covered by either of these conditions is that the X_k are *uniformly bounded*: $|X_k| \leq C$ for some C and all $k \geq 1$.)

In assessing this portion of Turing's paper, it is important to keep two points in mind. First, Turing states in his preface that "since reading Lindeberg's paper I have for obvious reasons made no alterations to that part of the paper which is similar to his." The manuscript is thus necessarily incomplete; it presumably would have been further polished and refined had Turing continued to work on it; the technical sufficient conditions given represent how far Turing had gotten on the problem *prior* to seeing Lindeberg's work. Second, in 1934 the Lindeberg condition was only known to be *sufficient*, not necessary; thus even in discussing his results in other sections of the paper (where he felt free to refer to the Lindeberg result), it may not have seemed important to Turing to contrast his own particular technical sufficient conditions with those of Lindeberg; the similarity in method must have seemed far more important.

3.4. One Counterexample. In Section 14 Turing concludes by giving a simple example of a sequence X_1, X_2, \dots that satisfies the quasi-necessary conditions (3.1), but not the CLT. For $n \geq 1$, let

$$P[X_n = \pm n] = \frac{1}{2n^2}; \quad P[X_n = 0] = 1 - \frac{1}{n^2}.$$

Then $E[X_n] = 0$, $\text{Var}[X_n] = E[X_n^2] = 1$, $s_n^2 = \text{Var}[S_n] = n \rightarrow \infty$, and $\rho_n^2 = 1/n \rightarrow 0$; thus (3.1) is satisfied. Turing then shows that if S_n/s_n converges, the limit distribution must have a discontinuity at zero, and therefore cannot be Gaussian.

It is interesting that Turing should happen to choose this particular example; although he does not note it, the sequence $\{S_n/s_n; n \geq 1\}$ has the property that $\text{Var}[S_n/s_n] \equiv 1$, but $\lim_{n \rightarrow \infty} S_n(\omega)/s_n = 0$ for almost all sample paths ω . This is an easy consequence of the first Borel-Cantelli lemma: because

$$\sum_{n=1}^{\infty} P[X_n \neq 0] = \sum_{n=1}^{\infty} \frac{1}{n^2} = \zeta(2) = \frac{\pi^2}{6} < \infty,$$

it follows that $P[X_n \neq 0 \text{ infinitely often}] = 0$; thus $P[\sup_n |S_n| < \infty] = 1$ and $P[\lim_{n \rightarrow \infty} S_n/s_n = 0] = 1$.

The existence of such sequences has an interesting consequence for the CLT. Let $\{Y_n: n \geq 1\}$ be a sequence of independent random variables, jointly independent of the sequence $\{X_n: n \geq 1\}$ and such that $P[Y_n = \pm 1] = \frac{1}{2}$. Let $T_n =: Y_1 + Y_2 + \cdots + Y_n$; then a trite calculation shows that $S_n + T_n$ satisfies the Feller condition (2.3), but not the Lindeberg condition (2.1). Let $t_n^2 =: \text{Var}[T_n]$; then $T_n/t_n \Rightarrow N(0, 1)$ and $\text{Var}[S_n + T_n] = s_n^2 + t_n^2$, hence

$$\begin{aligned} \frac{S_n + T_n}{\sqrt{\text{Var}[S_n + T_n]}} &= \frac{s_n}{\sqrt{s_n^2 + t_n^2}} \left(\frac{S_n}{s_n} \right) + \frac{t_n}{\sqrt{s_n^2 + t_n^2}} \left(\frac{T_n}{t_n} \right) \\ &= \left(\frac{1}{\sqrt{2}} \right) \left(\frac{S_n}{s_n} \right) + \left(\frac{1}{\sqrt{2}} \right) \left(\frac{T_n}{t_n} \right) \\ &\Rightarrow N(0, \tfrac{1}{2}). \end{aligned}$$

Thus the sequence $S_n + T_n$ does converge to a Gaussian distribution! This does not, however, contradict the Feller converse to the Lindeberg CLT; that result states that $S_n + T_n$, rescaled to have unit variance, cannot converge to the *standard* Gaussian $N(0, 1)$.

4. DISCUSSION. Turing's Fellowship Dissertation tells us something about Turing, something about the state of mathematical probability at Cambridge in the 1930s, and something about the general state of mathematical probability during that decade.

I. J. Good (1980, p. 34) has remarked that when Turing "attacked a problem he started from first principles, and he was hardly influenced by received opinion. This attitude gave depth and originality to his thinking, and also it helped him to choose important problems." This observation is nicely illustrated by Turing's work on the CLT. His dissertation is, viewed in context, a very impressive piece of work. Coming to the subject as an undergraduate, his knowledge of mathematical probability was apparently limited to some of the older textbooks such as "Czuber, Morgan Crofton, and others" (*Preface*, p. ii). Despite this, Turing immediately realized the importance of working at the level of distribution functions rather than densities; developed a method of attack similar to Lindeberg's; obtained useful sufficient conditions for convergence to the normal distribution; identified the conditions necessary for true central limit behavior to occur; understood the relevance of a Cramér-type factorization theorem in the derivation of such necessary conditions; and discovered the Feller subsequence phenomenon. If one realizes that the defects of the paper, such as they are, must largely reflect the fact that Turing had ceased to work on the main body of it after being apprised of Lindeberg's work, it is clear that Turing had penetrated almost immediately to the heart of a problem whose solution had long eluded many mathematicians far better versed in the subject than he. (It is interesting to note that Lindeberg was also a relative outsider to probability theory, and only began to work in the field a few years before 1922.)

The episode also illustrates the surprisingly backward state of mathematical probability in Cambridge at the time. Turing wrote to his mother in April, 1934: "I am sending some research I did last year to Czuber in Vienna [the author of several excellent German textbooks on mathematical probability], not having found anyone in Cambridge who is interested in it. I am afraid however that he may be dead, as he was writing books in 1891" (Hodges, 1983, p. 88). (Czuber had in fact died nearly a decade before, in 1925.)

This disinterest is particularly surprising in the case of G. H. Hardy, who was responsible for a number of important results in probabilistic number theory. But anyone who has studied the Hardy-Ramanujan proof of the distribution of prime divisors of an integer (1917), and compared it to Turán's (see Kac, 1959, pp. 71–74) will realize at once that the even most rudimentary ideas of modern probability must have been foreign to Hardy; see also Elliott (1979, pp. 1–5), Elliott (1980, pp. 16–20). Indeed, Paul Erdős believes that “had Hardy known the even least little bit of probability, with his amazing talent he would certainly have been able to prove the law of the iterated logarithm” (Diaconis, 1993). Perhaps this reflected in part the limited English literature on the subject. In 1927, when Harald Cramér visited England and mentioned to Hardy (his friend and former teacher) that he had become interested in probability theory, Hardy replied that “there was no mathematically satisfactory book in English on this subject, and encouraged me to write one” (Cramér, 1976, p. 516).

Finally, Turing's thesis illustrates the transitional nature of work in mathematical probability during the decade of the 1930s, before the impact of Kolmogorov's pioneering book *Grundbegriffe der Wahrscheinlichkeitsrechnung* (Kolmogorov, 1933) had been felt. In his paper Turing had thought it necessary to state and prove some of the most basic properties of distribution functions and their convolutions (in Sections 3 and 4, and Appendix C of the dissertation). His comment that his Appendix C “is only given for the sake of logical completeness and it is of little consequence whether it is original or not” (*Preface*, p. i), illustrates that such results, although “known,” did not enjoy general currency at the time. (It is all too easy to overlook today the important milestone in the literature of the subject marked by the publication in 1946 of Harald Cramér's important textbook *Mathematical Methods of Statistics*.)

It is also interesting to note Turing's approach to the problem in terms of convolutions of distribution functions rather than sums of independent random variables. Feller had similarly avoided the use of the language of random variables in his 1935 paper, formulating the problem instead in terms of convolutions. The reason, as Le Cam (1986, p. 87) notes, was that “Feller did not think that such concepts [as random variable] belonged in a mathematical framework. This was a common attitude in the mathematical community.”

Current mathematical attitudes towards probability have changed so markedly from the distrust and scepticism of earlier times that today the sheer magnitude of the shift is often unappreciated. Joseph Doob, whose own work dates back to this period, notes that “even as late as the 1930s it was not quite obvious to some probabilists, and it was certainly a matter of doubt to most nonprobabilists, that probability could be treated as a rigorous mathematical discipline. In fact it is clear from their publications that many probabilists were uneasy in their research until their problems were rephrased in what was then nonprobabilistic language” (Le Cam, 1986, pp. 93–94).

5. EPILOGUE: BLETCHLEY PARK. After his fellowship dissertation Turing “always looked out for any statistical aspects of [a] problem under consideration” (Britton, 1992, p. ix). This trait of Turing is particularly striking in the case of his cryptanalytic work during the second world war.

Turing left England for Princeton in 1936, to work with the logician Alonzo Church; he returned in 1938, after his Fellowship at King's College had been renewed. Recruited almost immediately by GC and CS (the Government Code and Cipher School), on September 4th, 1939 (one day after the outbreak of war) Turing

reported to Bletchley Park, the British cryptanalytic unit charged with breaking German codes, soon rising to a position of considerable importance. (Turing's work at Bletchley was the subject of a 1987 London play, "Breaking the Code," written by Hugh Whitmore and starring Derek Jacobi, of "I, Claudius" fame.)

The staff at Bletchley Park included many gifted people, distinguished in a number of different fields; among these were the mathematicians M. H. A. Newman, J. H. C. Whitehead, Philip Hall, Peter Hilton, Shaun Wylie, David Rees, and Gordon Welchman; the international chessmasters C. H. O'D. Alexander, P. S. Milner-Barry, and Harry Golombek; and others such as Donald Mitchie (today an important figure in artificial intelligence), Roy Jenkins (the later Chancellor of the Exchequer), and Peter Benenson (the founder of Amnesty International). Turing's chief statistical assistant in the later half of 1942 was another mathematician, I. J. Good, fresh from studies under Hardy and Besicovitch at Cambridge. (Good arrived at Bletchley on May 27, 1942, the day the *Bismarck* was sunk.) In recent years Good has written several papers (Good 1979, 1980, 1992, 1993a) discussing Turing's *ad hoc* development of Bayesian statistical methods at Bletchley to assist in the decrypting of German messages. (More general accounts of the work at Bletchley include Lewin, 1978, Welchman, 1982, and Hinsley and Stripp, 1993; see also the bibliography in Good, 1992.)

The specific details of Turing's statistical contributions are too complex to go into here. (Indeed, much of this information was until recently still classified and, perhaps for this reason, Good's initial papers on the subject do not even describe the specific cryptanalytic techniques developed by Turing; they give instead only a general idea of the type of statistical methods used. But in his most recent paper on this subject (Good, 1993a), Jack Good does provide a detailed picture of the various cryptanalytic techniques that Turing developed at Bletchley Park.) Three of Turing's most important statistical contributions were: (1) his discovery, independently of Wald, of some form of sequential analysis; (2) his anticipation of empirical Bayes methods (later further developed in the 1950s by Good and independently by Herbert Robbins); and (3) his use of logarithms of the Bayes factor (termed by Good the "weight of evidence") in the evaluation and execution of decryption. (For many references to the concept of weight of evidence, see, for example, Good, 1993b and the two indices of Good, 1983.) The units for the logarithms, base 10, were termed *bans* and *decibans*:

The reason for the name ban was that tens of thousand of sheets of paper were printed in the town of Banbury on which weights of evidence were entered in decibans for carrying out an important process called Banburismus . . . [Good, 1979, p. 394]

"Tens of thousands of sheets of paper . . ." This sentence makes it clear that Turing's contributions in this area were not mere idle academic speculation, but an integral part of the process of decryption employed at Bletchley.

One episode is particularly revealing as to the importance with which the Prime Minister, Winston Churchill, viewed the cryptanalytic work at Bletchley. On October 21, 1941, frustrated by bureaucratic inertia, Turing, Welchman, Alexander, and Milner-Barry wrote a letter *directly* to Churchill (headed "Secret and Confidential; Prime Minister only") complaining that inadequate personnel had been assigned to them; immediately upon its receipt Churchill sent a memo to his principal staff officer directing him to "make sure they have all they want on extreme priority and report to me that this had been done" (Hodges, 1983, pp. 219–221).

Much of I. J. Good's own work in statistics during the decades immediately after the end of the war was a natural outgrowth of his cryptanalytic work during it; this includes both his 1950 book *Probability and the Weighing of Evidence*; and his papers on the sampling of species (e.g., Good, 1953) and the estimation of probabilities in large sparse contingency tables (much of it summarized in Good, 1965). Some of this work was stimulated either directly (see, e.g., Good, 1973, p. 936) or indirectly (the influence being somewhat remote, however, in the case of contingency tables) by Turing's ideas:

Turing did not publish these war-time statistical ideas because, after the war, he was too busy working on the ground floor of computer science and artificial intelligence. I was impressed by the importance of his statistical ideas, for other applications, and developed and published some of them in various places. [Good, 1992, p. 211]

ACKNOWLEDGMENTS. I thank Anthony Edwards for his assistance in obtaining a copy of the typescript of Turing's Fellowship Dissertation during a visit to Cambridge in May 1992; and the Master and Fellows of Gonville and Caius College for their hospitality during that visit. Quotations from Turing's unpublished Fellowship Dissertation appear here by the kind permission of Professor Robin Gandy of Oxford University. Thanks also to Persi Diaconis, John Ewing, Jack Good, Steve Stigler, and an anonymous referee for helpful comments on an earlier draft of the paper.

REFERENCES

- Adams, W. J. (1974). *The Life and Times of the Central Limit Theorem*. Kaedmon, New York.
- Breiman, L. (1968). *Probability*. Addison-Wesley, Reading, MA.
- Britton, J. L., ed. (1992). *The Collected Works of A. M. Turing: Pure Mathematics*. North-Holland, Amsterdam. [Contains the two-page Preface to Turing's Fellowship Dissertation.]
- Burnside, W. (1928). *Theory of Probability*. Cambridge University Press.
- Cramér, H. (1936). Ueber eine Eigenschaft der normalen Verteilungsfunktion. *Mathematische Zeitschrift* 41, 405–414.
- Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press.
- Cramér, H. (1976). Half of a century of probability theory: some personal recollections. *Annals of Probability* 4, 509–546.
- Diaconis, P. (1993). Personal communication. [The quotation is a paraphrase from memory.]
- Diaconis, P. and Zabell, S. (1991). Closed form summation for classical distributions: variations on a theme of De Moivre. *Statistical Science* 6, 284–302.
- Elliott, P. D. T. A. (1979). *Probabilistic Number Theory I: Mean Value Theorems*. Springer-Verlag, New York.
- Elliott, P. D. T. A. (1980). *Probabilistic Number Theory II: Central Limit Theorems*. Springer-Verlag, New York.
- Feller, W. (1935). Über den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift* 40, 521–559.
- Feller, W. (1937). Über den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung, II. *Mathematische Zeitschrift* 42, 301–312.
- Feller, W. (1971). *An Introduction to Probability Theory and its Applications*, vol. 2, 2nd ed. Wiley, New York.
- Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika* 40, 237–264.
- Good, I. J. (1965). *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. M.I.T. Press, Cambridge, MA.
- Good, I. J. (1973). The joint probability generating function for run-lengths in regenerative binary Markov chains, with applications. *Annals of Statistics* 1, 933–939.
- Good, I. J. (1979). A. M. Turing's statistical work in World War II. *Biometrika*, 66, 393–396.
- Good, I. J. (1980). Pioneering work on computers at Bletchley. *A History of Computing in the Twentieth Century*, N. Metropolis, J. Howlett, and G.-C. Rota, eds. Academic Press, New York, pp. 31–45.
- Good, I. J. (1983). *Good Thinking*. Minnesota University Press.
- Good, I. J. (1992). Introductory remarks for the article in *Biometrika* 66 (1979). In *The Collected Works of A. M. Turing: Pure Mathematics* (J. L. Britton, ed.), North-Holland, Amsterdam, pp. 211–223.

- Good, I. J. (1993a). Enigma and Fish. In *Codebreakers: The Inside Story of Bletchley Park* (F. H. Hinsley and A. Stripp, eds.), Oxford University Press, pp. 149–166.
- Good, I. J. (1993b). Causal tendency, necessity and sufficiency: an updated review. In *Patrick Suppes, Scientific Philosopher* (P. Humphreys, ed.), Kluwer, Dordrecht (in press).
- Hardy, G. H. and Ramanujan, S. (1917). The normal number of prime factors of a number. *Quarterly J. Math.* 48, 76–92.
- Hodges, A. (1983). *Alan Turing: The Enigma*. Simon and Schuster, New York.
- Kac, M. (1959). *Statistical Independence in Probability, Analysis and Number Theory*. Carus Mathematical Monographs, Number 12. Mathematical Association of America.
- Kolmogorov, A. A. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Ergebnisse der Mathematik, Springer-Verlag, Berlin.
- Le Cam, L. (1986). The central limit theorem around 1935 (with discussion). *Statistical Science* 1, 78–96.
- Lévy, P. (1935). Propriétés asymptotiques des sommes de variables indépendantes on enchainées. *J. Math. Pures Appl.* 14, 347–402.
- Lévy, P. (1937). *Théorie de l'Addition des Variables Aléatoires*. Gauthier-Villars, Paris.
- Lewin, R. (1978). *Ultra Goes to War*. McGraw-Hill, New York.
- Lindeberg, J. W. (1922). Eine neue Herleitung des Exponential-gesetzes in der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift* 15, 211–225.
- Maistrov, L. E. (1974). *Probability Theory: A Historical Sketch*. Academic Press, New York.
- Petrov, V. V. (1975). *Sums of Independent Random Variables*. Springer-Verlag, New York.
- Stigler, S. M. (1986). *The History of Statistics*. Harvard University Press.
- Turing, A. M. (1934). On the Gaussian error function. Unpublished Fellowship Dissertation, King's College Library, Cambridge.
- Welchman, G. (1982). *The Hut Six Story*. McGraw-Hill, New York.

Departments of Mathematics and Statistics
Northwestern University
2006 Sheridan Road
Evanston, IL 60208
zabell@math.nwu.edu

How happy the lot of the mathematician. He is judged solely by his peers, and the standard is so high that no colleague or rival can ever win a reputation he does not deserve.

—W. H. Auden (1907–1973)
The Dyer's Hand, London: Faber & Faber, 1948, p. 15.

Niels Hendrik Abel and Equations of the Fifth Degree

Michael I. Rosen

This paper is dedicated to the memory of my close friend and colleague Kenneth Ireland.

In most textbooks it is stated that Abel was the first to prove that the general equation of the fifth degree cannot be solved in radicals. However, Abel's proof is almost never presented. Instead, the theorem is proved by means of Galois theory.

Abel published his first proof of this theorem (at his own expense) in 1824 [1, Vol 1], and a longer more elaborate version appeared in Crelle in 1826 [1, Vol. 1]. E. Galois was thirteen years old in 1824. His spectacular paper on the theory of equations was submitted to, and rejected by, the French Academy of Science in 1830. It wasn't published until 1846, fourteen years after his death. For details of this sad story the reader can consult the very interesting book of Harold Edwards [5]. From all this it is clear that Abel's proof could not have used Galois theory. How then did he do it? The purpose of this article is to provide an answer to this question which will be easily accessible to a modern reader familiar with the elements of the theory of fields. The proof we will give is not identical with that of Abel, but is in the spirit of his proof and uses nothing that was unavailable to him. Both before and after the proof we try to put things in historical context, and indicate how matters developed after 1826. In particular we will discuss the earlier work of P. Ruffini and the later work of Galois, as well as a pretty contribution of L. Kronecker.

Of course, other authors have discussed this material. What's new here is mainly the mode of presentation and the arrangement of the proofs. R. Ayoub's article on Ruffini [2] gives an excellent historical and mathematical overview of the theorem. J. P. Tignol's recent book on the theory of equations [7] gives among other things a history of the subject from ancient times up to the era of Galois. Both these sources discuss Abel's contributions. Nevertheless, we feel that a relatively brief and accessible exposition of these matters from a somewhat different point of view may be of interest to readers who are unfamiliar with this fascinating piece of mathematical history.

SECTION 1. The solution of the quadratic equation $x^2 + ax + b = 0$ goes back to antiquity. The roots are

$$x_1, x_2 = \frac{-a \pm \sqrt{a^2 - 4b}}{2}.$$

The solution of the cubic equation $x^3 + ax^2 + bx + c = 0$ was not discovered until the 16th century. Around 1515, S. del Ferro found a solution, but did not

publish it. The solution was rediscovered in 1535 by N. Fontana, nicknamed Tartaglia, who also kept it a secret until it was coaxed out of him by G. Cardano and published in Cardano's famous work "Ars Magna". The first step is to reduce the cubic $x^3 + ax^2 + bx + c$ to the form $x^3 + px + q$ by means of the substitution $x \rightarrow x - \frac{a}{3}$. The solutions of $x^3 + px + q = 0$ are given by

$$x_1, x_2, x_3 = \sqrt[3]{-\frac{q}{2} + \sqrt{\left(\frac{p}{3}\right)^3 + \left(\frac{q}{2}\right)^2}} + \sqrt[3]{-\frac{q}{2} - \sqrt{\left(\frac{p}{3}\right)^3 + \left(\frac{q}{2}\right)^2}}.$$

This formula must be supplemented by a rule telling how to choose the cube roots properly.

Soon after Tartaglia found his solution to the cubic equation, a solution was found to the quartic equation by L. Ferrari, the brilliant assistant to Cardano. We shall not write it down, but simply note that it involves nothing more than the rational operations of addition, subtraction, multiplication, and division, as well as extractions of square and cube roots. At this point it seemed reasonable to believe that the quintic equation could be solved by similar means, i.e. starting with the coefficients of the equations one should employ the rational operations together with the extraction of square, cube, and possibly fifth roots. However, in spite of much effort on the part of some of the best mathematicians in the world, no solution was found for over two and one half centuries.

The first mathematician to state definitively that no solution existed was Ruffini. In 1799, Ruffini published a two volume treatise entitled, "Teoria Generale delle Equazioni" in which he claims to show that the general equation of the fifth degree cannot be solved in radicals. For various reasons his results were received with skepticism, even though as eminent a mathematician as A. Cauchy found his arguments convincing. It turns out that although Ruffini did prove quite a lot, and did make important contributions, there was a significant gap in his proof. For all this the reader should consult [2]. We will discuss the gap in Ruffini's proof in Section 4.

While Ruffini's proof did not find universal acceptance, his work did help turn the direction of research away from the problem of finding a solution to an equation of the fifth degree to the problem of showing that in general no such solution exists. It is in this atmosphere that the young Abel entered the picture.

SECTION 2. In this section we set up notation, give a precise statement of the problem, review the solution via Galois theory, and begin our discussion of how Abel was able to find a proof which doesn't use Galois theory; a necessity for him since Galois theory had not yet been invented when he discovered his proof.

Throughout this paper we will use freely the notion of field, field extension, etc. Abel and his predecessors expressed themselves in different, but equivalent language. All fields will be assumed to be of characteristic zero.

Let k be a field and $f(x) \in k[x]$ a monic polynomial. If

$$f(x) = (x - \theta_1)(x - \theta_2) \cdots (x - \theta_n)$$

in some extension field of k , we call $F = k(\theta_1, \theta_2, \dots, \theta_n)$ a splitting field of $f(x)$ over k . In other words, a splitting field F of $f(x)$ over k is the field obtained from k by adjoining all the roots of $f(x) = 0$ to k .

A finite algebraic extension E/k is called a radical tower over k if there is a series of intermediate fields

$$k = E_0 \subset E_1 \subset \cdots \subset E_{m-1} \subset E_m = E$$

such that for each $0 \leq i \leq m$, $E_{i+1} = E_i(\sqrt[p_i]{\alpha_i})$ where p_i is a prime and $\alpha_i \in E_i^*$.

We can now give the precise definition of what it means for an equation to be solvable in radicals. Let $f(x) \in k[x]$ be a polynomial, and F a splitting field for $f(x)$ over k . We say that the equation $f(x) = 0$ is solvable in radicals if there is a radical tower E/k such that $F \subset E$. This definition is just a way of saying, in the language of fields, that the roots of $f(x) = 0$ can be obtained from the coefficients by the successive use of the rational operations and the extraction of roots. It is not clear, *a priori*, when $f(x) = 0$ is solvable in radicals, that F/k is itself a radical tower. In fact, this is not true in general as can be seen by considering the extension $\mathbf{Q}(\alpha)/\mathbf{Q}$ where $\alpha = 2 \cos(2\pi/7)$. This extension is not a radical extension. On the other hand, α is a root of the irreducible cubic $x^3 + x^2 - 2x - 1$ which splits into linear factors in $\mathbf{Q}(\alpha)$.

One of the principal accomplishments of Galois was to give a beautiful criterion for when an equation $f(x) = 0$ was solvable in radicals. It is no real loss of generality to assume that $f(x)$ is irreducible, and we do so. In this circumstance, Galois shows how to assign a group G_f to $f(x)$. It is a certain transitive subgroup of the group of permutations of the roots of $f(x)$.

Theorem (Galois). $f(x) = 0$ is solvable in radicals if and only if G_f is a solvable group.

We recall that a finite group G is solvable if there is a sequence of subgroups

$$(e) = G_0 \subset G_1 \subset G_2 \subset \cdots \subset G_m = G$$

such that for each i , $0 \leq i < m$, G_i is normal in G_{i+1} and $p_{i+1} = [G_{i+1} : G_i]$ is prime.

To use Galois' theorem to show equations of the fifth degree and higher are not in general solvable in radicals, one computes the Galois group of the general equation of the n th degree and shows it is equal to S_n , the full symmetric group on n letters. One then shows that S_n is not a solvable group when $n \geq 5$. This is the approach used in all modern texts in algebra.

Let's explain the notion of the "general equation of degree n ". Let k be a field of characteristic zero, and let s_1, s_2, \dots, s_n be quantities which are algebraically independent over k . Set $K = k(s_1, s_2, \dots, s_n)$ and define

$$f(x) = x^n - s_1 x^{n-1} + s_2 x^{n-2} - \cdots + (-1)^n s_n \in K[x]$$

to be the general equation of degree n over k .

Suppose $f(x) = (x - x_1)(x - x_2) \cdots (x - x_n)$ in some extension field of K . Set $L = K(x_1, x_2, \dots, x_n)$. Clearly, L is a splitting field for $f(x)$ over K . It is not hard to show that x_1, x_2, \dots, x_n are algebraically independent over k (for details, see [6]). Moreover, the s_i are elementary symmetric functions of the x_i .

$$\begin{aligned} s_1 &= x_1 + x_2 + \cdots + x_n \\ s_2 &= x_1 x_2 + x_1 x_3 + \cdots + x_{n-1} x_n \\ &\vdots \\ s_n &= x_1 x_2 \cdots x_n. \end{aligned}$$

Each permutation of the x_i induces an automorphism of L which leaves K fixed. Moreover, the only elements of L which are left fixed by all such automorphisms are the elements of K . Although this is stated in modern language, the content is all quite old. The last part is easily seen from the theorem that a symmetric polynomial is a polynomial in the elementary symmetric functions of the variables. This fact goes back, in essence, to Newton (see [5]), and was used freely by the predecessors of Abel and Galois. Reverting to modern language, we see that L/K is a Galois extension with Galois group isomorphic to S_n , or, put another way, S_n is the Galois group of the general equation of degree n over k . So, with this setup a certain amount of Galois theory was available to people like Vandermonde, Lagrange, Ruffini, and Abel. What was missing was the notion of normal subgroup. This fundamental notion is not visible in the work of these earlier authors. Thus they could not even formulate the notion of a solvable group, never mind prove Galois' criterion for when an equation was solvable in radicals. The rudiments of group theory will play a big role in our treatment of Abel's work on this problem, but nowhere will the notion of normal subgroup make an appearance.

SECTION 3. We will now devote our attention to properties of the group S_n . Very little will be needed. The elements of S_n will be considered to be permutations of the set $\{1, 2, \dots, n\}$. For a polynomial in n variables $f(x_1, x_2, \dots, x_n)$ and an element $\sigma \in S_n$ we define

$$(\sigma f)(x_1, \dots, x_n) = f(x_{\sigma(1)}, \dots, x_{\sigma(n)}).$$

This action extends in a natural way to an action on rational functions. Now define

$$\delta = \prod_{i < j} (x_i - x_j) \quad \text{and} \quad \Delta = \delta^2.$$

For any $\sigma \in S_n$, $\sigma\delta = \pm\delta$. If τ is a transposition, $\tau\delta = -\delta$. Then, $A_n \subset S_n$ is defined to be the subset of S_n consisting of all $\sigma \in S_n$ such that $\sigma\delta = \delta$. Clearly, A_n is a subgroup. Note that $[S_n : A_n] = 2$ since for any $\sigma \in S$ and any transposition τ , either σ or $\tau\sigma$ is in A_n .

Facts:

1. S_n is generated by transpositions.
2. A_n is generated by 3-cycles.
3. A_n is generated by m -cycles, where m is any odd number between 3 and n .

The first two facts are standard. To prove the third, note that an m -cycle is a product of $m - 1$ transpositions, so if m is odd an m -cycle is in A_n . On the other hand, the identity

$$(a_1 a_2 a_3) = (a_2 a_1 a_3 a_4 \dots a_m)(a_m a_{m-1} \dots a_4 a_3 a_2 a_1)$$

shows that every 3-cycle is in the group generated by the m -cycles, so Fact 3 follows from Fact 2.

These facts are all we will need, but we add one more, due to Cauchy, since Abel made use of it in his original proof.

4. Let S_n act on $L = K(x_1, x_2, \dots, x_n)$ as explained above. Let p be the largest prime less than or equal to n . Then, for $f \in L$, the number of distinct values taken on by f under the action of S_n (i.e. the number of distinct rational functions obtained from f by permuting the variables) either exceeds p or is at most 2.

Proof: Let $\sigma \in S_n$ be a p -cycle, and $\langle \sigma \rangle$ be the subgroup generated by σ . Define $H = \{\tau \in \langle \sigma \rangle \mid \tau f = f\}$. Since p is a prime, either $H = \langle \sigma \rangle$ or $H = \langle e \rangle$. Thus, either $f, \sigma f, \dots, \sigma^{p-1}f$ are all distinct, or $\sigma f = f$. If f takes on fewer than p values, we must have $\sigma f = f$ for all p -cycles. By Fact 3 this implies that f is fixed by A_n . Since A_n has index 2 in S_n the result follows.

It is, of course, true that this result is really about S_n acting on an arbitrary set (same proof), but we have given the original formulation.

SECTION 4. We have now assembled everything we shall need. We use the notation of Section 2, except that we now add the assumption, as Abel did, that sufficiently many roots of unity are in the ground field k . Readers who are bothered by this can take $k = \mathbb{C}$, the complex numbers. Nothing essential is lost by this.

Theorem (Abel). *Let $f(x) = x^n - s_1 x^{n-1} + \dots + (-1)^n s_n$ be the general equation of degree n over k . If $n \geq 5$ then this equation is not solvable in radicals.*

Recall that $f(x) = (x - x_1)(x - x_2) \dots (x - x_n)$ in $L = K(x_1, x_2, \dots, x_n)$. S_n acts on L by permuting the x_i and $K = k(s_1, s_2, \dots, s_n)$ is the fixed field. If $f(x) = 0$ were solvable in radicals, we would have a radical tower E/K such that $L \subseteq E$. Abel proceeds in two steps.

Step 1. If L is contained in a radical tower over K , then L/K is itself a radical tower.

Step 2. If $n \geq 5$ then L/K is not a radical tower (in fact, Abel restricts himself to the case $n = 5$).

When he discovered his proof, Abel was unaware that the proof of Step 2 had been achieved years earlier by Ruffini. Ruffini did not give a proof of Step 1. It is not clear that he realized it was necessary. That was the gap in his proof! So, the proof of Step 1 was Abel's essential contribution.

In the next section we will give a proof of Step 2 by a method different from that of either Abel or Ruffini, although it is close in spirit to some of Ruffini's later proofs (he gave many). I have adapted it from the classic text of Burnside and Panton [3]. The proof is short and elegant and uses nothing that was unavailable to either Abel or Ruffini. In Section 6 a proof of Step 1 will be given which is in essence that of Abel except for the use of the language of field theory and the inclusion of more details than are given in the original paper. At the end of that section we will also sketch a portion of Abel's own proof of Step 1.

SECTION 5. With the previous notation still in effect, we will show that L/K cannot be a radical tower if $n \geq 5$.

We re-emphasize that we are assuming that the base field k contains as many roots of unity as needed.

Suppose

$$K = K_0 \subset K_1 \subset K_2 \subset \dots \subset K_n = L$$

is a radical tower. Then there is a prime p and an element $a \in K^*$ such that $K_1 = K(\sqrt[p]{a})$. We will show that $p = 2$ and that $a = b^2 \Delta$ where $b \in K^*$ and Δ is

the symmetric function defined at the beginning of Section 3. In other words, K_1 is uniquely determined and is the field $K(\sqrt[p]{\Delta})$.

To prove this, set $\alpha = \sqrt[p]{a}$ and let $\tau \in S_n$ be a transposition. Applying τ to $\alpha^p = a$ we find $\tau(\alpha)^p = a$ and consequently, $(\tau(\alpha)/\alpha)^p = 1$ so that $\tau(\alpha) = \zeta\alpha$ where $\zeta^p = 1$. Now apply τ to both sides of the equation $\tau(\alpha) = \zeta\alpha$ and one finds $\alpha = \tau(\zeta\alpha) = \zeta\tau(\alpha) = \zeta^2\alpha$. It follows that either $\tau(\alpha) \neq \alpha$ for some transposition τ and $p = 2$, or α is fixed by all transpositions. In the latter case, α is fixed by all S_n (use Fact 1 of Section 3) and is an element of K , contrary to assumption. The proof shows that $\tau(\alpha) = \pm\alpha$ for all transpositions τ . It follows that $\sigma(\alpha) = \pm\alpha$ for all $\sigma \in S_n$. Now, every three-cycle is a square; in fact, $(abc) = (acb)^2$. Thus, α is fixed by three-cycles and so by all of A_n by Fact 2 of Section 3. Since $\tau(\alpha) = -\alpha$ for at least one transposition we now see this must be true for every transposition. This property also holds for the polynomial δ . Consequently, α/δ is fixed by all transpositions and so also by all elements in S_n . Thus, $\alpha/\delta = b \in K$ and so $a = \alpha^2 = b^2\delta^2 = b^2\Delta$ which shows $K_1 = K(\sqrt{\Delta})$ as asserted.

We will show that when $n \geq 5$, K_1 has no radical extension in L . This will complete the proof that L/K is not a radical tower when $n \geq 5$.

Suppose $c \in K_1^*$ and $K_2 = K_1(\sqrt[q]{c})$. Set $\gamma = \sqrt[q]{c}$. By the first part of the proof, A_n leaves K_1 fixed. Let ρ be a three-cycle, and apply ρ to both sides of the equation $\gamma^q = c$. One deduces that $\rho(\gamma) = \zeta\gamma$ where $\zeta^q = 1$. Applying ρ twice to the equation $\rho(\gamma) = \zeta\gamma$ yields $\gamma = \rho^3(\gamma) = \zeta^3\gamma$. Thus, either $\rho(\gamma) = \gamma$ for all three-cycles ρ or $\rho(\gamma) \neq \gamma$ for some three-cycle and $q = 3$. In the former case, γ is fixed by A_n and is in K_1 , contrary to assumption. As we point out below, one need not invoke Galois theory to establish this point. The conclusion is that $q = 3$. If $n \geq 5$, A_n is also generated by five-cycles (use Fact 3 of Section 3). Repeating the same arguments shows that $q = 5$. This contradiction establishes the result.

Remark 1. The fact that $K_1 = K(\sqrt{\Delta})$ is the fixed field of A_n can be proven as follows. Let $\gamma \in L$ be fixed by A_n and let τ be a transposition. It is not hard to see that $a = \gamma + \tau(\gamma)$ and $b = \gamma \cdot \tau(\gamma)$ are both fixed by S_n and so are in K . But, γ is a root of the equation $x^2 - ax + b = 0$ and so must generate a quadratic extension of K in L . We have seen that $K_1 = K(\sqrt{\Delta})$ is the unique quadratic extension of K in L , and so $\gamma \in K_1$.

Remark 2. The group theory behind the proof is very simple and general. Let G be a group generated by elements of order r and of order s . Suppose that $\gcd(r, s) = 1$. Then, G has no abelian quotients. For, if G/N is an abelian quotient, it must be generated by elements of order r and elements of order s . Since it is an abelian group it is annihilated by r and s , and so also by 1 which is the gcd of r and s . Thus, $G = N$. In the above proof we took $G = A_n$ for $n \geq 5$, and $r = 3$, $s = 5$.

SECTION 6. We now come to Abel's proof of Step 1; if L/K is contained in a radical tower it is a radical tower. We continue to assume that $K = k(s_1, s_2, \dots, s_n)$ and that L is the splitting field of the generic polynomial $x^n - s_1x^{n-1} + \dots + (-1)^ns_n$. However, the arguments are quite general and apply whenever L/K is the splitting field of a separable polynomial over K , and K contains sufficiently many roots of unity.

All the results we will need are in Abel's original article [1, pp. 66–87] with perhaps different formulations. The first Lemma is now a standard result.

Lemma 1. Let F be a field containing a primitive q 'th root of unity. If $a \in F^*$ is not a q 'th power, the $x^q - a$ is irreducible.

If α is a root of $x^q - a = 0$ then every $\gamma \in F(\alpha)$ can be written in the form

$$\gamma = a_0 + a_1\alpha + \cdots + a_{q-1}\alpha^{q-1} \quad (1)$$

where the a_i are in F .

Lemma 2. Assume that $x^q - a \in F[x]$ is irreducible and that α is a root. Let γ be an element of $F(\alpha)$ with $\gamma \notin F$. Then there is a $\beta \in F(\alpha)$ such that $\beta^q \in F$ and

$$\gamma = b_0 + \beta + b_2\beta^2 + \cdots + b_{q-1}\beta^{q-1}$$

where $b_0, b_2, \dots, b_{q-1} \in F$.

Proof: Write γ as in equation (1) above. Let $1 \leq k < q$ be the smallest integer such that $a_k \neq 0$. Set $\beta = a_k\alpha^k$. Clearly, $\beta^q \in F$. For $1 \leq m < q$ we can find integers r and s such that $0 \leq s < q$ and $rq + sk = m$. Then

$$\alpha^m = (\alpha^q)^r (\alpha^k)^s = c_s \beta^s \quad \text{with } c_s \in F.$$

The desired expression for γ now follows by substitution into equation (1).

Lemma 3. Let q be a prime, and ζ a primitive q 'th root of unity. Then, for each integer i ,

$$1 + \zeta^i + \zeta^{2i} + \cdots + \zeta^{(q-1)i} = \begin{cases} 0 & \text{if } q \text{ does not divide } i, \\ q & \text{if } q \text{ divides } i. \end{cases}$$

Proof: Again, this is standard. If q divides i the result is clear. If q doesn't divide i one uses the formula for the sum of a geometric series.

Lemma 4. Consider the extension L/K . Let $y \in L$. Then the irreducible polynomial for y over K splits into linear factors in $L[x]$.

Proof: Let y_1, y_2, \dots, y_m be the distinct values (conjugates) of y under the action of the symmetric group. Then, $g(x) = (x - y_1)(x - y_2) \cdots (x - y_m)$ has coefficients which are invariant under S_n and so are elements of K . The irreducible polynomial for y over K must divide $g(x)$ and the result follows. (Of course, it is easy to see that the irreducible polynomial for y over K is $g(x)$).

We now come to the main lemma which contains the crux of the argument. Roughly speaking, it asserts that if a radical extension containing K is intersected with L , the resulting pair of fields is again a radical extension. It might be worthwhile at this point to remind the reader once more that we are assuming all the roots of unity that arise are in the base field. If this assumption is not made, the result may well be false.

Lemma 5. Let E/K be an extension field, q a prime, and $a \in E$ an element such that $x^q - a \in E[x]$ is irreducible. Let α be a root of $x^q - a = 0$. Set $M = E(\alpha) \cap L$ and $M_0 = E \cap L$. If $M \neq M_0$ then M/M_0 is a radical extension. More precisely, there is a $\beta \in M$ such that $\beta^q \in M_0$ and β generates M over M_0 .

Proof: Let $y \in M$, $y \notin M_0$. By Lemma 2, we can find a $\beta \in E(\alpha)$ such that $\beta^q = b \in E$ and

$$y = b_0 + \beta + b_2\beta^2 + \cdots + b_{q-1}\beta^{q-1}$$

where the $b_i \in E$. Let $g(x) \in K[x]$ be the irreducible polynomial for y over K , and set

$$G(x) = g(b_0 + x + b_2x^2 + \cdots + b_{q-1}x^{q-1}).$$

$G(x)$ is in $E[x]$ and has β for a root. By Lemma 1, $x^q - b \in E[x]$ is irreducible. Thus $x^q - b$ divides $G(x)$. It follows that $G(\zeta^i\beta) = 0$ where ζ is a primitive q 'th root of unity, and i is any integer, and so the numbers

$$\begin{aligned} y &= y_1 = b_0 + \beta + b_2\beta^2 + \cdots + b_{q-1}\beta^{q-1} \\ y_2 &= b_0 + \zeta\beta + b_2\zeta^2\beta^2 + \cdots + b_{q-1}\zeta^{q-1}\beta^{q-1} \\ &\vdots \\ y_q &= b_0 + \zeta^{q-1}\beta + b_2\zeta^{2(q-1)}\beta^2 + \cdots + b_{q-1}\zeta^{(q-1)(q-1)}\beta^{q-1} \end{aligned} \quad (2)$$

are all roots of $g(x)$. By Lemma 4, we see that the numbers y_1, y_2, \dots, y_q are all in L (we implicitly assume that L and $E(\alpha)$ are contained in some common extension field). Multiply the i 'th equation in (2) by ζ^{1-i} and add all the resulting equations. Using Lemma 3, we find

$$\beta = \frac{1}{q} \sum_{i=1}^q \zeta^{1-i} y_i \in L.$$

Thus $\beta \in L \cap E(\alpha) = M$ and $\beta^q = b \in L \cap E = M_0$.

It remains to show that β generates M over M_0 . Let $\gamma \in M$. We can write

$$\gamma = c_0 + c_1\beta + c_2\beta^2 + \cdots + c_{q-1}\beta^{q-1} \quad c_i \in E.$$

It will suffice to show that the coefficients $c_i \in E \cap L = M_0$. To do this one repeats the above argument to show that

$$\gamma_i = c_0 + c_1\zeta^{i-1}\beta + \cdots + \zeta^{(i-1)(q-1)}\beta^{q-1}$$

is in $L \cap E(\alpha) = M$ for $i = 1, 2, \dots, q$. Multiply γ_i by $\zeta^{k(1-i)}$ and add up over i ranging from 1 to q . Using Lemma 3 once more we find

$$c_k\beta^k = \sum_{i=1}^q \zeta^{k(1-i)}\gamma_i \in M.$$

Since $\beta \in M$, it follows that $c_k \in M \cap E = M_0$ which completes the proof.

This argument is so pretty and ingenious, one is lost in admiration! We are now ready to state and prove the main result.

Theorem. *If L/K is contained in a radical tower, then L/K is itself a radical tower.*

Proof: Suppose that E/K is a radical tower and that $L \subseteq E$. We have

$$K = E_0 \subset E_1 \subset E_2 \subset \cdots \subset E_m = E$$

where $E_{i+1} = E_i(\sqrt[q_i]{a_i})$, q_i being a prime, and $a_i \in E_i$.

Now, consider the tower

$$K = E_0 \cap L \subseteq E_1 \cap L \subseteq \cdots \subseteq E_{m-1} \cap L \subseteq L. \quad (3)$$

If $E_{i+1} \cap L = E_i \cap L$ there is nothing that need be said. If $E_{i+1} \cap L \neq E_i \cap L$ then Lemma 5 shows that $E_{i+1} \cap L / E_i \cap L$ is a radical extension (of degree q_i). Thus, after eliminating equalities, equation (3) demonstrates L as a radical tower over K .

This completes the proof of Step 1 of Section 4. Since we proved Step 2 in the last section, the proof that the general equation of degree 5 or greater cannot be solved in radicals is now complete.

Abel's original proof of Step 2 is different from the one we have given, and we want to give an idea of how he did it. To do this we sketch Abel's proof that any radical extension of K inside L has degree 2. We will assume, as Abel did, that we are dealing with the general equation of degree 5.

Let $K \subset K_1 \subset L$ and suppose $K_1 = K(\alpha)$ where $\alpha^q = a \in K$ and q is a prime. Let m be the number of distinct values that α takes on under the action of S_5 (i.e. the number of distinct conjugates of α). As we have seen, m is the degree of the irreducible equation for α over K . Since $x^q - a$ is irreducible it follows that $m = q$. Since $|S_5| = 120$, q must divide 120, i.e. $q = 2, 3$, or 5 . By Cauchy's result, Fact 4 of Section 3, q cannot equal 3. Thus $q = 2$ or $q = 5$, and we must show $q = 5$ is impossible. Abel does this by first showing that the only fields between K and L of degree 5 over K are the fields $K(x_i)$, where $i = 1, 2, 3, 4, 5$. If $q = 5$ we can then assume that $K(\alpha) = K(x_1)$. By modifying α , if necessary, as in the proof of Lemma 2, we can write

$$x_1 = a_0 + \alpha + a_2\alpha^2 + a_3\alpha^3 + a_4\alpha^4 \quad a_i \in K.$$

Applying the same technique as in Lemma 5 one shows

$$\alpha = \frac{1}{5}(x_1 + \zeta^{-1}x_2 + \zeta^{-2}x_3 + \zeta^{-3}x_4 + \zeta^{-4}x_5) \quad (4)$$

where ζ is a primitive fifth root of unity. The contradiction arises from the fact that under the action of S_5 , α has five values, but the right hand side of (4) has 120 values.

Abel goes on to show that $K_1 = K(\sqrt[5]{\Delta})$ and that $K(\sqrt[5]{\Delta})$ has no radical extensions in L . The proof of the latter assertion is similar to the one we have just given. If I am not mistaken there is a minor flaw in Abel's proof that $K_1 = K(\sqrt[5]{\Delta})$, but this is very minor and is easily corrected.

SECTION 7. We conclude by indicating some further developments. In this section we continue to assume that we are in characteristic zero, but will no longer demand that roots of unity be in the base field.

Abel was fascinated with the theory of equations. He published three articles on the subject, and a fourth appears among his posthumous work (see item XVIII of Volume II of his collected works [1]). He was at work on a major new memoir on this theory when he died, tragically, at the early age of 27.

Having proved that the general equation of degree 5 or greater cannot be solved in radicals, the thrust of his later work was to find conditions on special equations which insure that they can be solved in radicals. His best known result in this direction is the following proposition.

Proposition 1 (Abel). *Let $f(x) \in k[x]$ and suppose that $\theta_1, \theta_2, \dots, \theta_n$ are its roots in some extension field of k . Suppose each θ_i is a rational function of θ_1 , i.e. each $\theta_i = R_i(\theta_1)$ where $R_i(x) \in k(x)$. Suppose further that for each pair i and j we have*

$$R_i(R_j(\theta_1)) = R_j(R_i(\theta_1)).$$

Then $f(x) = 0$ is solvable in radicals.

The reader can see that the hypotheses can be translated as follows. The splitting field of $f(x)$ is generated by θ_1 , and the Galois group of $f(x)$ is abelian. Thus the proposition is a consequence of Galois theory, though this is not, of

course, how Abel proved it. It is probably because of this result that Abel's name is attached to groups in which the elements commute with one another.

Abel never published a general criterion for when an equation is solvable in radicals, but in a letter to Crelle dated October 18, 1828 (see [1], Vol. 2), he writes

“Si trois racines d'une équation quelconque irréductible dont le degré est un nombre premier, sont liées entre elles de sorte que l'une de ces racines puisse être exprimée rationnellement par le deux autres, l'équation en question sera toujours résoluble à l'aide de radicaux.”

Roughly translated, this reads “If every three roots of an irreducible equation of prime degree are related to one another in such a way that one of them may be expressed rationally in terms of the other two, then the equation is solvable in radicals”. Abel gives no indication of how he came to this result, or how he proved it. It is remarkable in part because the statement is almost identical to one of the principal results of Galois' fundamental memoir of 1830, which as we have already pointed out was not published until 1846. Here is Galois' statement of the result as translated into English by Harold Edwards in [5].

Proposition 2 (Galois). *In order for an irreducible equation of prime degree to be solvable in radicals it is necessary and sufficient that once any two of the roots are known, that the others can be deduced from them rationally.*

One can rephrase this result in more modern language. Let $f(x) \in k[x]$ be irreducible of prime degree, and $\theta_1, \theta_2, \dots, \theta_p$ be its roots. Given any three roots $\theta_i, \theta_j, \theta_m$ there exists a rational function $R(x, y) \in k[x, y]$ such that $\theta_m = R(\theta_i, \theta_j)$. Or, more simply, the splitting field of $f(x)$ is generated by any two of its roots.

Considering the simplicity and beauty of this result, it is somewhat surprising that it is not better known. A proof is outlined in the exercises to Section 8, chapter 4 of [6]. A complete proof can be found in Section 5 of Chapter 14 of [7]. Edwards [5] also gives a complete proof, but since he uses Galois' original language it is somewhat hard to read.

It might be objected that this criterion is not useful because the hypothesis is very difficult to check. As it turns out, it can be quite useful. In 1856, L. Kronecker proved the following interesting result (see [4]).

Proposition 3 (Kronecker). *Let \mathbf{Q} be the rational numbers, and suppose $f(x) \in \mathbf{Q}[x]$ is an irreducible polynomial of prime degree. If $f(x) = 0$ is solvable in radicals, then either $f(x)$ has exactly one real root, or all its roots are real.*

Kronecker's proof uses the methods of Abel. He was clearly unaware of Galois' work, since this proposition is an immediate corollary of Proposition 2. If θ_1 and θ_2 are any pair of real roots, and $R(x, y) \in \mathbf{Q}[x, y]$, Then clearly $R(\theta_1, \theta_2)$ is also real, and so all the roots must be real.

It is easy to use Proposition 2 to produce polynomials in $\mathbf{Q}[x]$ which are not solvable in radicals. For example, let $q \geq 5$ and p be primes, and $a \geq 2$ be an integer. Let $f(x) = x^q - apx - p$. By Eisenstein's criterion, $f(x)$ is irreducible. We claim it has exactly three real roots. For x large and negative, $f(x)$ is negative. At $x = -1$, $f(-1) = -1 + p(a - 1) > 0$. At $x = 0$, $f(0) = -p < 0$. Finally, when x is large and positive, $f(x)$ is positive. By the intermediate value theorem, $f(x)$ has at least three real roots. However, $f'(x) = qx^{q-1} - ap$ has exactly two real roots, so $f(x)$ must have exactly three real roots. By Kronecker's result it

follows that $f(x) = 0$ cannot be solved in radicals. The simplest special case is $x^5 - 4x - 2$.

It is of some interest to point out that a l -adic version of Proposition 3 is valid.

Proposition 3l. *Suppose $f(x) \in \mathbb{Q}[x]$ is of prime degree. Let \mathbb{Q}_l denote the field of l -adic numbers. If $f(x) = 0$ is solvable in radicals, then either exactly one root of $f(x)$ is in \mathbb{Q}_l or all its roots are in \mathbb{Q}_l .*

As before, the proof is an immediate consequence of Proposition 2.

Here is an example of how to use this. Consider the polynomial

$$f(x) = x^5 + 3x^4 + 3x^3 + 6x^2 + 3x + 6.$$

$f(x)$ is an Eisenstein polynomial at 3 and so it is irreducible over \mathbb{Q} . Considering $f(x)$ modulo 2 we find

$$f(x) \equiv x^5 + x^4 + x^3 + x \equiv x(x+1)(x^3+x+1) \pmod{2}.$$

Since $f(x)$ has exactly two roots modulo 2, both of which are simple roots, one can invoke Hensel's lemma to conclude that $f(x)$ has exactly two roots in \mathbb{Q}_2 . It follows from Proposition 3l that $f(x) = 0$ cannot be solved in radicals.

We conclude by using these ideas to answer the question; do there exist polynomials in $\mathbb{Q}[x]$ of prime degree, all of whose roots are real, but which are not solvable in radicals? Here is an amusing way to show that the answer is yes. Let p and l be primes. Let $f_1(x)$, $f_2(x)$, and $f_3(x) \in \mathbb{Q}[x]$ be polynomials of the same prime degree $q \geq 5$ such that $f_1(x)$ is an Eisenstein polynomial at p , $f_2(x)$ has exactly two distinct roots modulo l , and $f_3(x)$ has q distinct real roots. Use the weak approximation theorem to find a polynomial $f(x) \in \mathbb{Q}[x]$ which is p -adically close to $f_1(x)$, l -adically close to $f_2(x)$, and close in archimedean absolute value to $f_3(x)$. Then, since $f(x)$ is p -adically close to $f_1(x)$ it is irreducible. Since it is close to $f_3(x)$ in the archimedean absolute value, it has all its roots real. Finally, since it is l -adically close to $f_2(x)$, it has exactly two distinct roots in \mathbb{Q}_l and by Proposition 3l it is not solvable in radicals.

All of this provides a nice, if simple, example of the fruitful way old and new mathematics can be combined to good effect.

REFERENCES

1. N. H. Abel, *Oeuvres Complètes*, Two Volumes, (L. Sylow and S. Lie, ed.), Grondahl and Son, Christiana, 1881.
2. R. Ayoub, Paolo Ruffini's Contributions to the Quintic, *Arch. Hist. Exact Sci.*, Vol. 23, pp. 253–277, 1980.
3. W. S. Burnside and A. W. Panton, *The Theory of Equations*, Vol. 2, Longmans, Green, and Co., London-New York-Toronto, 1928.
4. H. Dörrie, *One Hundred Great Problems of Elementary Mathematics*, Dover Publ., New York, 1965.
5. H. Edwards, *Galois Theory*, Springer Verlag, New York-Berlin-Heidelberg-Tokyo, 1984.
6. N. Jacobson, *Basic Algebra 1*, W. H. Freeman and Co., San Francisco, 1974.
7. J.-P. Tignol, *Galois' theory of algebraic equations*, Longman Scientific Technical co-published with John Wiley and Sons, New York, 1987.

Mathematics Department
Brown University
Providence, RI 02912
MA408000@brownvm.bitnet

The Great Marble Race: An Assignment Gone Wrong

Benny Evans and Jerry Johnson

As many educators have discovered, it is not always easy to produce computer-related activities that are accessible to students and actually contribute to their understanding of mathematics. Occasionally the most carefully designed assignments in the hands of students produce unexpected results. This is not necessarily bad—on the contrary, it may be a learning opportunity. The purpose of this note is to discuss an interesting example of this.

Before we start, we will say a word about software. We use the computer algebra system *DERIVE*® in our classes and what follows reflects this fact. There are several other such programs on the market, most notably Maple™ and Mathematica®, with which the exercise we describe below can be analyzed as well. The most important thing is the mathematics, not the software, but users of other programs may observe some differences in the results. As we will see, the essence of the problem we discuss is structural, so any software should ultimately lead to the same difficulty.

BACKGROUND. In 1990 the authors produced a book of computer-based activities entitled *Uses of Technology in the Mathematics Curriculum*. Written under NSF grant USE 8950044, the book features substantial laboratory exercises whose solutions require a computer algebra system or other computational aid. They were the end result of our own experience, as well as that gained from participants at an NSF Workshop we held in 1989. This planning made us confident that we had anticipated most responses our students were likely to give... but not all. Here's the story of what happened with one of these problems.

The following is part of an exercise from our book that we assigned to students in Calculus II. The complete version actually begins with a “warm-up” which is a straightforward solved problem to provide a context for the computer commands necessary for the solution of the full exercise. We added necessary *DERIVE* instructions and we took some liberty—to be precise, the correct model is a frictionless bead sliding down a ramp, since a rolling ball picks up rotational energy. For a nice discussion and analysis of the Brachistochrone problem see [3].

MLRC ASSIGNMENT. MATH 2265 THE GREAT MARBLE RACE: THE BRACHISTOCHRONE

This is a marble race. We start our marbles at the origin and let them roll down a decreasing path to the point $(\pi, -2)$. (We have chosen this particular point to

DERIVE is a registered trademark of Soft Warehouse, Inc.

make our results nicer.) The trick is, we each get to select the path we want the marble to follow. I choose the Brachistochrone, given by $(\pi t - \sin \pi t, \cos \pi t - 1)$, $0 \leq t \leq 1$. Someone else has already chosen the straight line, $(\pi t, -2t)$, $0 \leq t \leq 1$, and a third contestant has chosen the parabolic arc $(\pi t, 2(t - 1)^2 - 2)$, $0 \leq t \leq 1$. You may select any curve you want. Just be sure your parametric curve decreases from the origin to $(\pi, -2)$ as t moves from 0 to 1.

1. (a) What path do you wish to use for the marble race?
(b) For the path you have chosen, show that $(x(0), y(0)) = (0, 0)$ and $(x(1), y(1)) = (\pi, -2)$.
2. To decide who wins, we need to calculate the time required for a marble to roll down a path given by $(x(t), y(t))$, $0 \leq t \leq 1$, from the origin to the point $(\pi, -2)$. The following formula is derived by equating the increase in kinetic energy gained by the marble with the potential energy it loses in getting to a spot on the curve a distance y below the x -axis.

$$\text{Time} = \frac{1}{\sqrt{g}} \int_0^1 \sqrt{\frac{(x'(t))^2 + (y'(t))^2}{-2y(t)}} dt$$

where g is the constant of gravitation attraction near the surface of the earth.

3. For each of the following curves, provide an integral that gives the time to finish. You should evaluate the integrals for the brachistochrone and the straight line by hand, then ask *DERIVE* to **approximate** the answer. *DERIVE* cannot evaluate the integral for the parabola in closed form. (If you issue the **Simplify** command, *DERIVE* will try for a while and then give up.) It is likely to have a similar difficulty with your curve and you will have to ask *DERIVE* to **approximate** again. In each approximation, you may leave the gravitational constant g as a letter. For example, your answer may look like $(1 + g)5.12345$. Fill in the following table:

curve	integral	exact time	time to 6 places
Brachistochrone			
straight line			
parabola			
your curve			

4. Who won the race?
5. Based on the outcome of the marble race, formulate a conjecture about a physical property of the Brachistochrone. (Only clear English sentences are acceptable here.)

NOTE: The full exercise continues with an analysis of the tautochrone property of the curve.

THE PHILOSOPHY BEHIND THE EXERCISE. We liked this problem for several reasons. For one, the answer frequently runs counter to the intuition of many students that a straight ramp will win the race. It also encourages experimentation. Students can compute the time integral for different paths using *DERIVE* to do

the calculations, which would be prohibitive otherwise. The proof that the Brachistochrone is the path of least time is pretty challenging for calculus students, but this gives them the opportunity to test a few paths to find evidence for believing it.

STUDENT RESPONSE. This is not an easy problem, and some students had difficulty coming up with a parametric representation of a path that decreases from the origin to $(\pi, -2)$. Since we provided linear and quadratic paths we anticipated that they would probably choose a cubic path because it is the next polynomial degree.

The first semester we gave the exercise, all the students did what we expected and came up with the corresponding times for these paths. The better students actually tried to justify the brachistochrone's victory by giving some nice heuristic arguments about how the marble should accelerate quickly at the outset. The poor students produced the same kind of work they would without a computer. One made such a mess of the integral that it turned out to be a complex number, which *DERIVE* correctly calculated. The student dutifully reported the finish time as $3 + 4i$.

THE ASSIGNMENT GONE WRONG. The next semester, a student tried something that got him very excited. He came to Professor Johnson's office to announce that he had chosen the path $(\pi t, -2\sqrt{t})$ and obtained a finish time of 3.08830 which is less than the Brachistochrone time of π . (*Remark:* These answers should be divided by \sqrt{g} , but we are only interested in a comparison, so we have suppressed this constant here and in all answers that follow.)

We examined the student's work for an error, but there was none. The student actually began to wonder out loud if he had uncovered an age-old mistake. At this point we cranked up *DERIVE* and took a look!

The time integral cannot be evaluated in closed form and must be approximated; moreover the time integral is improper (or singular) in the sense that the integrand approaches infinity as t approaches 0 (see Figure 4). *DERIVE* signals its uncertainty about the accuracy of an approximation by beeping and displaying the phrase "dubious accuracy" on the screen. The accuracy is indeed dubious, and the problem became clear.

We changed *DERIVE*'s precision setting from the default of 6 digits to 10, and after a few seconds *DERIVE* produced an approximation of 3.154709288, a number that would look better to Bernoulli. (In the next section we'll discuss why this was a futile thing to do.) This calculation took 30.5 seconds on our 386 33MHz machine, whereas the six digit approximation took 3.2 seconds. Nevertheless, we could simply instruct students to set the precision to 10 digits before beginning the exercise and warn them to be patient. *But*, another clever student used a segment of the ellipse $(\pi t, -2\sqrt{2t - t^2})$. With 10 digit precision the ellipse gives a value of 3.132361048 in 37.3 seconds, making the ellipse a winner over the Brachistochrone even with a 10 digit setting.

It seems that the idea behind the exercise has been defeated, and the teacher may be placed in the awkward position of pronouncing the Brachistochrone to be the winner in spite of the evidence that students have gathered to the contrary.

WHERE DO WE GO FROM HERE? Is this all *DERIVE*'s fault for providing poor approximations? As we will see, this is not the case. The "dubious accuracy" message is evidence of a structural impediment and increasing the precision settings, as we did above, is likely to fail.

```

1:  [w t, - 2 √(- t (t - 2))]
2:  [w t - SIN (w t), COS (w t) - 1]
3:  [w t, - 2 √t]
4:  [w t, 2 t (t - 2)]
5:  "Four ramps from (0,0) to (π,-2)."
```

```

COMMAND: Author Build Calculus Declare Expand Factor Help Jump solve Manage
          Options Plot Quit Remove Simplify Transfer move Window approx
Enter option
User      C:B3.MTH      Free:96% Insert      Derive Algebra
```

Figure 1. Four parametric paths.

A look at Figure 1 shows a *DERIVE* screen with four paths. #1 is the second student's ellipse mentioned above, #2 is the Brachistochrone, #3 is the first student's parabola, and #4 is the parabola we suggested in the exercise. Figure 2 shows the graphs numbered accordingly (note the scale at the bottom).

Why the difference among them in the accuracy of the corresponding time integrals? To help understand, look at Figure 3 which displays the integrands of the time integrals for the four paths. Figure 4 shows the graphs of these integrands numbered to correspond to the four graphs in Figure 2.

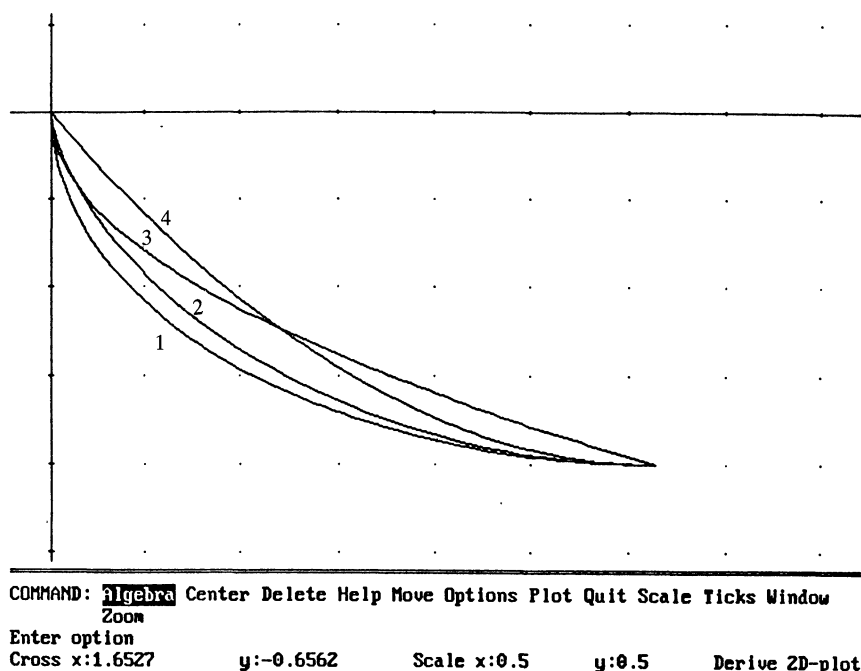


Figure 2. Graphs of the four paths.

We can see that three of the integrals are singular: the integrands approach ∞ as $x \rightarrow 0$. The approximation techniques rely on a finite number of function values, so if the function changes dramatically over any of the corresponding subintervals, accuracy is uncertain. The rapid growth of the function near the

60:

$$\sqrt{\frac{\frac{1}{2} t (t - 2) - 4 (t^2 - 2 t + 1)}{(- t (t - 2))^{3/2}}}$$

61:

62:

$$\sqrt{\frac{\frac{1}{2} t + 1}{t^{3/2}}}$$

63:

$$\sqrt{\frac{1}{2} t^2 + 16 t^2 - 32 t + 16} \sqrt{-\frac{1}{t (t - 2)}}$$

COMMAND: Author Build Calculus Declare Expand Factor Help Jump solve Manage Options Plot Quit Remove Simplify Transfer move Window approx

Enter option

User C:BRACH.MTH Free:71% Insert Derive Algebra

Figure 3. The *integrands* of the time integrals of the four paths.

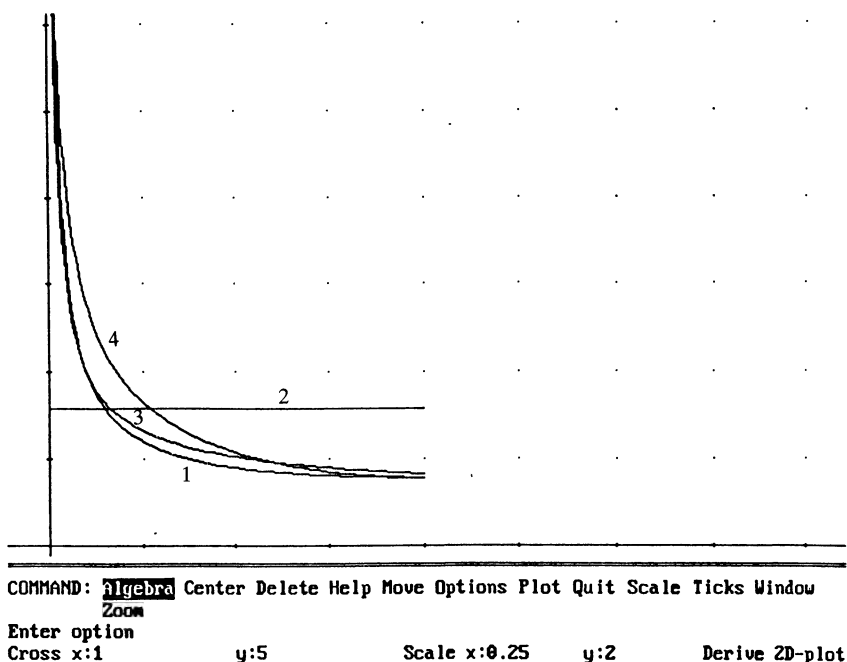


Figure 4. Graphs of the *integrands* of the time integrals of the four paths.

singularity at the origin or the occurrence of a large percentage of the area under the curve near the y -axis contribute to this difficulty and make results of “dubious accuracy” more likely. A look at the cases where the two student answers were less than π (#1 and #3 in Figures, 1, 2, and 4) shows that this is happening.

The computing time for most arithmetic operations grows quadratically with the precision. Increasing *DERIVE*’s precision enables more refinement which further

increases the computation time and memory consumption. This can improve the approximation for a given function, but it is easy to produce functions that will defeat any precision setting. See [2] for more on these matters.

We were encouraged by students who chose curves that drop very steeply and then quickly flatten out, giving a large acceleration at the origin. However, this forces the time integral to get a large part of its value near the singularity. The more students try to increase acceleration in this way, the more trouble the computer algebra system is likely to have approximating the time integral.

We can also make this situation as bad as we want by changing parameterizations. Let's look at the parabolic arc #4 in Figure 1 and consider the parameterizations $(\pi t^r, 2(1 - t^r)^2 - 2)$ where r is a small positive number. The graphs of the *integrands* of the time integral for $r = 0.1, 0.2, \dots, 1$ are displayed in Figure 5, and descend as r decreases. Once again, please note that the curves in Figure 5 are *not* the ramps that our marble rolls down, but the graphs of the functions we integrate to get the time of descent. These integrals all represent the time down the same parabolic ramp and so must have the same value, but *DERIVE* produces dramatically different approximations (with “dubious accuracy” messages where appropriate).

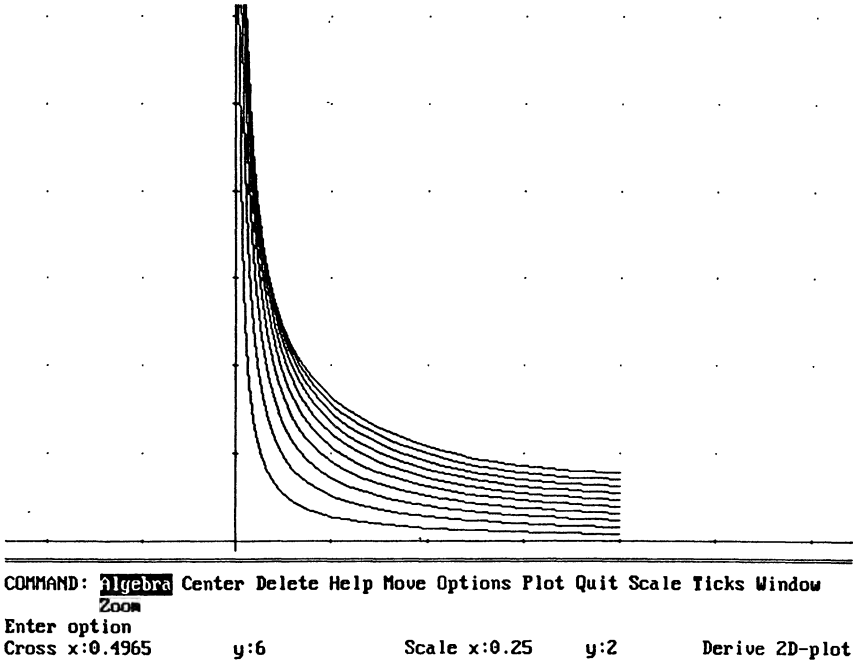


Figure 5. Integrands for different parameterizations of a parabolic arc.

As r gets smaller the integral takes more of its value from a region near the singularity, $x = 0$. Thus, an approximation determined by a regular partition will become successively poorer. Using the default 6-digit precision, *DERIVE* produces an approximation of 2.98912 for $r = 0.4$ and 1.41820 for $r = 0.1$ with “dubious accuracy” warnings. In the latter case the time is less than half the Brachistochrone’s. Thus, the idea of increasing *DERIVE*’s precision is defeated by a sufficiently small choice of r .

For Riemann integrals, there are ways to determine bounds on the error in an approximation, such as those given by Simpson's rule and the Trapezoidal rule, but if we ask our computer algebra system to approximate an *improper* integral, we have no way to assure the accuracy of the answer. This is not the fault of the computer algebra system. No matter what general approximation technique we employ, there is a fairly straightforward integral that will defeat it. See [2] for more on these matters. What are we to do?

RESOLVING THE PROBLEM: GIVE YOUR COMPUTER ALGEBRA SYSTEM SOME HELP. There are techniques that can help approximate improper integrals in general. First, one should determine that the integral converges before turning it over to a computer algebra system. If the integral diverges, there is no point in appealing to the computer.

If the integral extends over an infinite interval, a substitution can be made to change it to a finite interval. For example,

$$\int_0^\infty f(x) dx = \int_0^1 \frac{f(-\ln x)}{x} dx \quad \text{or} \quad \int_1^\infty f(x) dx = p \int_0^1 \frac{f(x^{-p})}{x^{p+1}} dx.$$

(*DERIVE* does this automatically, but if we leave the process up to our computer algebra system, we have no control over what the transformed integral looks like.)

Once we have an improper integral over a finite interval our best strategy is to transform it into a Riemann integral if that is possible. For a large class of functions this is easy. For example, if $f(t)$ is continuous on $(0, 1]$ such that $\int_0^1 f(t) dt$ converges and $\lim_{t \rightarrow 0} t^r f(t)$ is finite for some r with $0 \leq r < 1$ (which is true for algebraic functions), then the substitution $t = x^{1/1-r}$ will transform the improper integral into a Riemann integral.

By the way, in the case of our time integral, making a substitution in the integral is the same as changing the parameterization of the path—a good exercise for a calculus student. We see this in the next example.

Consider the parameterization of the parabolic arc $(\pi t^{1/10}, 2(1 - t^{1/10})^2 - 2)$. This yields the time integral

$$\int_0^1 f(t) dt \quad \text{where} \quad f(t) = \frac{1}{20} \sqrt{\frac{\pi^2 + 16(t^{1/5} - 2t^{1/10} + 1)}{t^{19/10}(t^{1/10} - 2)}}.$$

As was noted earlier, *DERIVE* provides a “dubious accuracy” warning and the unsatisfactory approximation 1.41820 for this integral. It is important to emphasize that if we did not know in advance to expect the answer to be larger than π we might have no idea how bad this approximation really is. Since $\lim_{t \rightarrow 0} t^{19/20} f(t)$ is finite we should make the substitution $t = x^{20}$ to yield the Riemann integral

$$\int_0^1 \sqrt{\frac{\pi^2 + 16(x^4 - 2x^2 + 1)}{2 - x^2}} dx$$

for which *DERIVE* gives the approximation 3.27633. It is also instructive to plot the graph of the integrand, which suggests that the integral is larger than π .

This trick may also work for transcendental functions. For example, using its 6-digit precision, *DERIVE* provides the approximation 0.886195 for $\int_0^1 \sqrt{-\ln t} dt$. Notice that $\lim_{t \rightarrow 0} t^{1/2} \sqrt{-\ln t}$ is finite, and hence the substitution $t = x^2$ produces the integral $\int_0^1 2\sqrt{2}x\sqrt{-\ln x} dx$ for which *DERIVE* gives 0.886239. As it turns out, the original estimate was not too bad, but we had no way of knowing this before we transformed the integral.

Of course there are integrals that do not satisfy the above criterion. For example, if $f(t) = t^{-1}e^{-\sqrt{-\ln t}}$ then $\int_0^1 f(t) dt$ converges to 2 (*DERIVE* will find the antiderivative in closed form), but $\lim_{t \rightarrow 0^+} t^r f(t) = \infty$ for all $0 < r < 1$. (We leave the proof to the reader.) The substitution $t = x^r$ for a large value of r may still be advisable since it shifts area under the integrand away from the singularity, but in general there may be no recipe to tell us how to transform such improper integrals into Riemann integrals.

THE ERROR. Approximations without error bounds are not worth much. Thus, while *DERIVE*'s approximations for the transformed Riemann integrals were improved in the examples we examined, we can't be sure of this in general. According to its User Manual, *DERIVE* employs "an extrapolated adaptive Simpson's rule" to approximate Riemann integrals. What this means is that *DERIVE* adjusts the distribution of partition points in its application of Simpson's rule. For more discussion of this method, see [1]. Simpson's rule has a standard error bound in terms of the maximum of the absolute value of the fourth derivative on the interval in question. In the two examples discussed above the derivative tends to ∞ at one of the end points. Thus, we appear to be back to square one. Do we have reason to accept *any* of the approximations that *DERIVE* has produced?

In general, this is a serious problem that has no easy answer, but in special cases there is help. If $f(x)$ is monotone on the interval $[a, b]$, and if the interval is partitioned into n equal subintervals, then the difference between the upper and lower Riemann sums is $|f(b) - f(a)|(b - a)/n$, so the average of the upper and lower Riemann sums (the trapezoidal rule) differs from the integral by no more than $|f(b) - f(a)|(b - a)/2n$.

Therefore, in simple cases we may apply the trapezoidal rule on intervals where the function is monotone. It is easy to calculate the required Riemann sums with

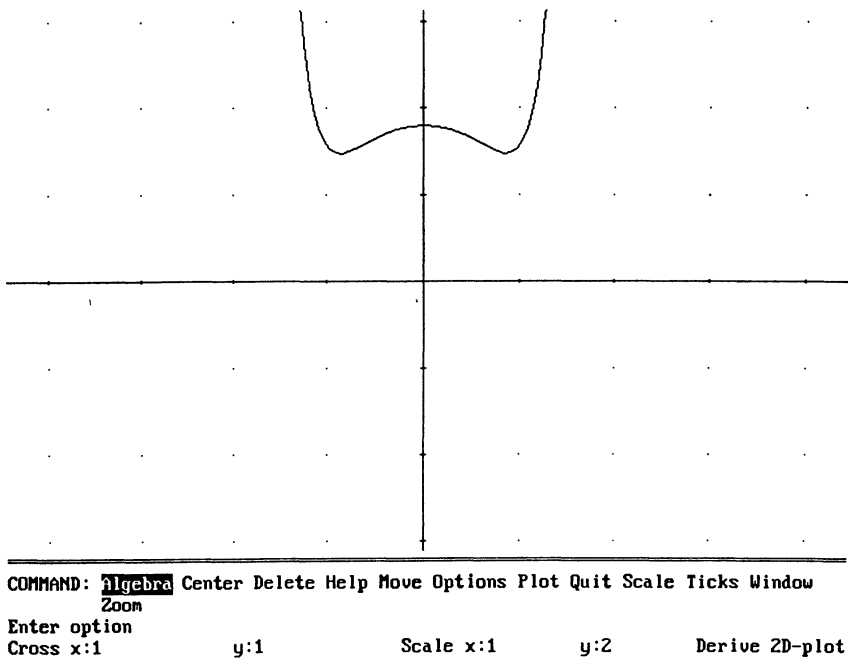


Figure 6. The graph of $\sqrt{\frac{\pi^2 + 16(x^4 - 2x^2 + 1)}{2 - x^2}}$

DERIVE and to decide what value of n ensures the desired error bound. Let us look once again at the parameterization of the parabolic arc discussed above. With *DERIVE*'s help we find that the function

$$\sqrt{\frac{\pi^2 + 16(x^4 - 2x^2 + 1)}{2 - x^2}}$$

has a single extremum on the interval $[0, 1]$ at $x = 0.853490$. If we wish to get an answer with error less than 10^{-3} we apply the trapezoidal rule on the interval $[0, 0.853490]$ with $n = 1300$ and on the interval $[0.853490, 1]$ with $n = 400$. *DERIVE* produces the value $2.83576 + 0.440535 = 3.27630$ and we finally have a number whose accuracy is understood.

CONCLUSION. This problem provides a natural setting where an improper (or singular) integral arises and a context for such questions as “What are the difficulties involved in approximating such integrals?” “How does the parameterization affect the integral?” “How can you help the computer produce estimates that you can believe?” When we discussed these topics in class, we found an interested audience because some of them had seen the difficulties.

In this example students see that even the mighty computer does not provide an easy solution to every calculation, and it is not facility with software or formulas, but rather the ability to bring appropriate mathematics to bear that will solve problems. In particular, it shows that computation may lead one astray and that in the end, a mathematical *proof* is the essential tool.

The unforeseen problem with this exercise was turned into a positive learning experience which probably makes it better, not worse, than we believed when we assigned it. The things that a modern computer algebra system can do are striking, and they provide mathematics educators with exciting tools that simply have never been available before, but it may be that the things a computer algebra system *cannot* do are just as important for pedagogical reasons.

ACKNOWLEDGMENT. We would like to thank David Stoutemyer, co-author of the *DERIVE* program, for examining the manuscript and making several helpful suggestions.

REFERENCES

1. Conte and de Boor, *Elementary Numerical Analysis*, McGraw-Hill, New York.
2. Rice, *Numerical Methods, Software and Analysis*, McGraw-Hill, 1983, New York.
3. Simmons, *Calculus with Analytic Geometry*, McGraw-Hill, 1985, New York.

Department of Mathematics
Oklahoma State University
Stillwater, OK 74078

Department of Mathematics
University of Nevada, Reno
Reno, NV 89557
jjohnson@math.unr.edu

Geometry and the Foucault Pendulum

John Oprea

Nature uses only the longest thread to weave her patterns, so each small piece of fabric reveals the organization of the entire tapestry.

—Richard P. Feynman

§1. INTRODUCTION. In 1851 Jean Foucault (1819–1868) built a pendulum consisting of a heavy iron ball on a wire 200 feet long to demonstrate the rotation of the Earth (see Figure 1a and Figure 1b). Foucault observed that such rotation would cause the swing-plane of the pendulum to precess, or rotate, as time went on, eventually returning to its original direction after a period of $T = 24/\sin \nu_0$ hours (where ν_0 denotes the latitude where the experiment takes place).

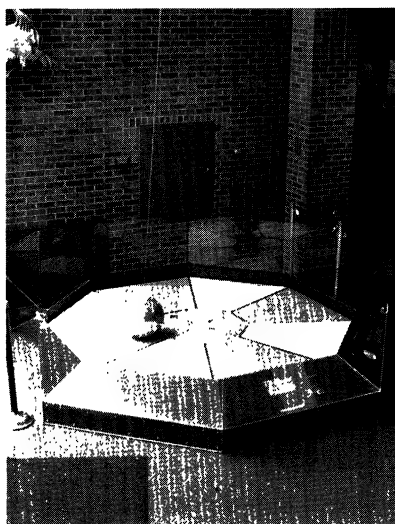


Figure 1a

In a recent *New York Times* interview [Ang], the distinguished scientist and author Stephen Jay Gould proclaimed, “I’ve never understood why every science museum in the country feels compelled to have one of these [a Foucault pendulum]. I still don’t understand how they work and I don’t think most visitors do either.” Gould is exactly right. Non-physicists generally have only the vaguest notion of how the behavior of the pendulum relates to the rotation of the Earth. The usual quite complicated analysis of this phenomenon of precession is in terms of rotating reference frames and the Coriolis force (see [Sym] and [Arn]). While these notions are part of elementary mechanics, they are not widely known among even mathematically aware non-physics students.

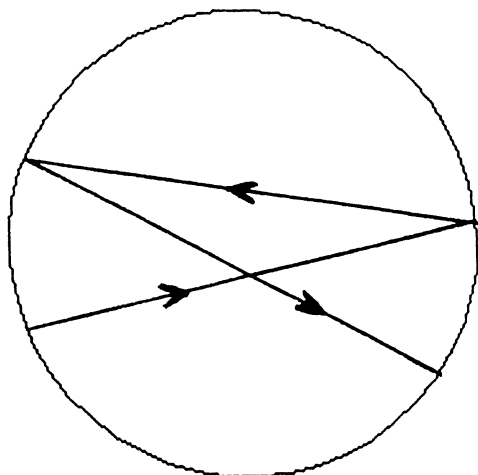


Figure 1b. Path of the Pendulum

The purpose of this article is to present the behavior of the Foucault pendulum as a simple consequence of doing Calculus on the sphere. This *holonomy* approach to the pendulum is mentioned in [W-S] and [Mar p. 16], but the details *in terms of elementary Calculus* do not seem to be well known. We believe this analysis of the Pendulum deserves a wide audience because it provides a beautiful down-to-'Earth' example of mathematical modelling in the context of Geometry and Calculus.

While we only discuss the pendulum, the geometric concept of holonomy makes its presence felt in applied mathematics from optimal control to quantum mechanics (cf. [En1], [En2] and [W-S]). It is hoped that the mathematical description of the Foucault pendulum presented here will spur interest in applications of Differential Geometry and will be accessible to any student acquainted with multivariable calculus and a touch of linear algebra.

§2. THE SPHERE. Our first step in analyzing Foucault's pendulum is to understand the geometry of the sphere. Consider a sphere (denoted by S^2) of radius R with patch

$$x(u, v) = (R \cos u \cos v, R \sin u \cos v, R \sin v),$$

where $0 \leq u \leq 2\pi$ and $-\frac{\pi}{2} \leq v \leq \frac{\pi}{2}$. By 'patch' we mean a system of coordinates on the sphere, such as spherical coordinates (ρ, θ, ϕ) with a fixed radius $\rho = R$. Note however that our patch differs from spherical coordinates in that v represents the latitude on the sphere; that is, the angle *up* from the equator, *not down* from the North Pole (see Figure 2).

The patch x has two special families of curves associated to it: the *longitudes* $\beta(v) = x(u_0, v)$ obtained by setting u equal to a constant and the *latitudes* $\alpha(u) = x(u, v_0)$ obtained by setting v equal to a constant. Since these curves are in \mathbb{R}^3 , their tangent vectors α' and β' are given by differentiating each coordinate of their expressions. For latitude and longitude tangent vectors respectively, we have

$$\alpha' = (-R \sin u \cos v_0, R \cos u \cos v_0, 0),$$

$$\beta' = (-R \cos u_0 \sin v, -R \sin u_0 \sin v, R \cos v).$$

Note that the dot product $\alpha' \cdot \beta'$ is zero, so that α' and β' are perpendicular (or

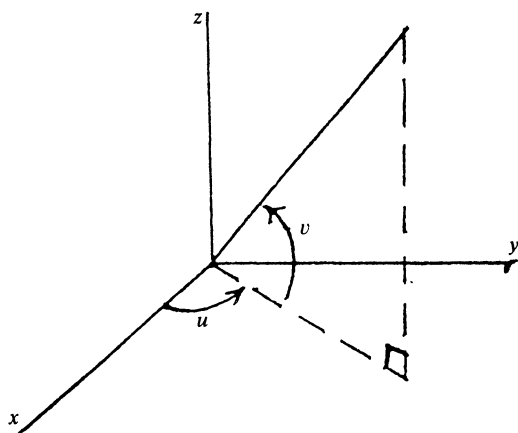


Figure 2

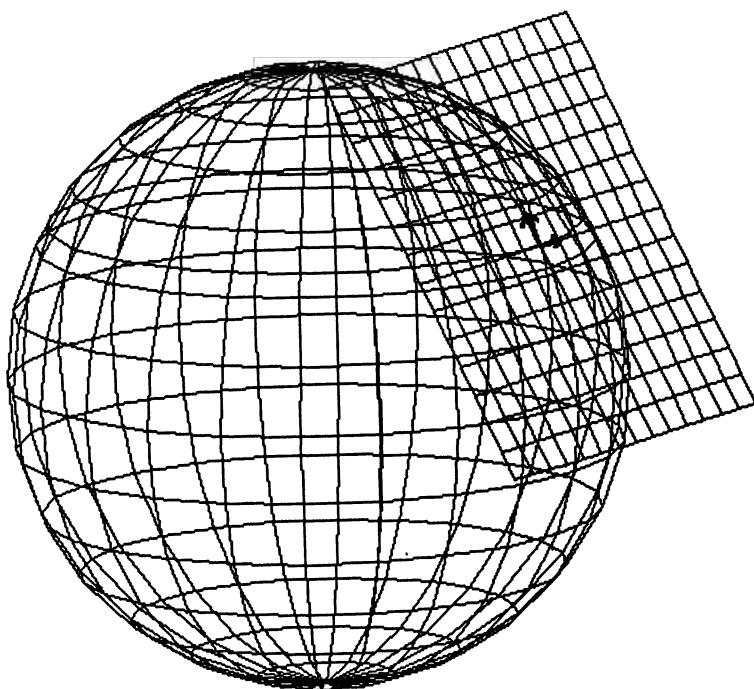


Figure 3. Tangent Plane with Basis E_1, E_2

orthogonal) for all u and v . In particular, α' and β' form a *basis* for the tangent plane $T_p S^2$ where $p = x(u_0, v_0)$. That is, every tangent vector w at $x(u, v)$ may be written in a unique way as $w = A\alpha' + B\beta'$ for some real numbers A and B (see Figure 3).

This basis for the tangent plane may be extended to a basis for \mathbb{R}^3 itself by taking a vector perpendicular to both α' and β' ; namely, the cross product $\alpha' \times \beta'$.

In fact, things become simpler if we take unit vectors in the directions of α' , β' and $\alpha' \times \beta'$ obtained by dividing these vectors by their lengths $|\alpha'|$, $|\beta'|$ and $|\alpha' \times \beta'|$. The vectors of our basis are now,

$$E_1 = \frac{\alpha'}{|\alpha'|} = (-\sin u, \cos u, 0) \quad E_2 = \frac{\beta'}{|\beta'|} = (-\cos u \sin v, -\sin u \sin v, \cos v)$$

and

$$U = \frac{\alpha' \times \beta'}{|\alpha' \times \beta'|} = (\cos u \cos v, \sin u \cos v, \sin v).$$

The basis $\{E_1, E_2, U\}$ provides a framework for comparing Euclidean geometry of \mathbb{R}^3 to geometry seen from the perspective of a 2-dimensional resident of the sphere. Because the perceptions of such a person are restricted to the 2-dimensional space spanned by E_1 and E_2 , any event or object in \mathbb{R}^3 is ‘seen’ by the resident of the sphere only through its projection onto the tangent plane. In particular, a vector w in \mathbb{R}^3 may be written uniquely as

$$w = aE_1 + bE_2 + cU$$

but the resident of the sphere only sees $aE_1 + bE_2$. The viewpoint described here is useful in forming analogies between Euclidean geometry and curved geometry. For example, in \mathbb{R}^3 we know that lines, which may be parametrized by $\gamma(t) = p + tv$ for fixed p and v , are shortest paths between points. Further, from the parametrization, it is clear that lines are characterized by having zero acceleration vectors. By analogy, ‘shortest paths’ (or *geodesics*) on the sphere are characterized by having zero acceleration vectors *as perceived by residents of the sphere*. That is, any curve on the sphere with an acceleration vector entirely in the U -direction is a geodesic. Such curves on the sphere turn out to be the great circles. In the next section we carry this viewpoint further.

§3. PARALLEL VECTORS ON THE SPHERE. What does it mean to say that two tangent vectors on the sphere in different tangent planes are parallel? It definitely cannot mean, in general, that the two vectors are parallel in \mathbb{R}^3 . For consider a latitude circle on the sphere S^2 at latitude v_0

$$\alpha(u) = (R \cos u \cos v_0, R \sin u \cos v_0, R \sin v_0).$$

It is easy to compute that, in \mathbb{R}^3 , $\alpha'(0)$ may be written as

$$\alpha'(0) = -R \sin v_0 \cos v_0 E_2 \left(\frac{\pi}{2}, v_0 \right) + R \cos^2 v_0 U \left(\frac{\pi}{2}, v_0 \right)$$

with respect to the basis $\{E_1, E_2, U\}$ at $\alpha(\frac{\pi}{2})$. The non-zero U -component shows that no vector of the tangent plane at $\alpha(\frac{\pi}{2})$ is \mathbb{R}^3 -parallel to $\alpha'(0)$.

One way to compare vectors along a curve $\gamma(t)$ in \mathbb{R}^3 is to start with a tangent vector V_0 at $\gamma(0)$ and create a *field* of tangent vectors $V(t)$ at $\gamma(t)$ which is differentiable in t . The rate of change in vectors along γ may then be computed as $(d/dt)V(t)$. Further, we may say that a vector field V is *parallel along* γ if $(d/dt)V(t) = 0$ for all t . Of course this then implies that $V(t) = V_0$, a constant, and this fits with our notion of parallelism in \mathbb{R}^3 .

We may extend this idea in a simple way to a tangent vector field $V(u)$ along a latitude circle $\alpha(u)$ in S^2 by saying that V is *parallel along* α if $(d/du)V(u)$ has no $E_1(u)$ or $E_2(u)$ components. This means that $(d/du)V(u) = C(u)U(u)$ for all u or, equivalently, that the projection of $(d/du)V(u)$ onto the tangent plane at $\alpha(u)$,

$\text{proj}_{TS^2}(d/du)V(u)$, is zero. We may think of this as saying that residents of the sphere see no change in vectors along α . (For readers versed in differential geometry, note that we may avoid the covariant derivative here because α is a constant-length u -parameter curve and $V(u)$ is given in terms of u . Thus, covariant differentiation in \mathbb{R}^3 , which is coordinatewise directional differentiation, reduces to ordinary differentiation d/du .)

To return to our latitude circle, let $V(u)$ be a parallel vector field along the latitude $\alpha(u)$. (We always assume that vectors are tangent to S^2 .) Then we may write $V(u) = A(u)E_1(u) + B(u)E_2(u)$. The first thing we notice is

Lemma. V has constant length.

Proof: Because V is parallel, $(d/du)V(u) = C(u)U(u)$ and therefore,

$$\begin{aligned}\frac{d}{du}(V(u) \cdot V(u)) &= 2\frac{d}{du}V(u) \cdot V(u) \\ &= C(u)U(u) \cdot V(u) \\ &= 0.\end{aligned}$$

Since $V \cdot V$ is constant, so is $|V|$. □

From our expression for $V(u)$ we see that we must have $A(u)^2 + B(u)^2 = |V|^2 = L^2$ where L is a constant. Therefore we may write $A(u) = L \cos \theta(u)$, $B(u) = L \sin \theta(u)$ where $\theta(u)$ is the angle from $V(u)$ to $E_1(u)$. We then have

$$V(u) = L \cos \theta(u)E_1(u) + L \sin \theta(u)E_2(u).$$

From this expression it is clear that, in order to compute $(d/du)V(u)$, we must first compute $(d/du)E_1(u)$ and $(d/du)E_2(u)$. We do this coordinatewise.

$$\frac{d}{du}E_1 = (-\cos u, -\sin u, 0) \quad \frac{d}{du}E_2 = (\sin u \sin v_0, -\cos u \sin v_0, 0).$$

The reader may check that, in terms of the basis $\{E_1, E_2, U\}$ we have

Proposition.

$$\frac{d}{du}E_1 = \sin v_0 E_2 - \cos v_0 U, \quad \frac{d}{du}E_2 = -\sin v_0 E_1.$$

Remark. Note that the Proposition says that neither E_1 nor E_2 are parallel along α .

The second thing we notice is that parallel vector fields always exist. In fact, the proof of this standard (but essential) result tells us precisely how vectors rotate to maintain parallelism.

Theorem. Let V_0 be a tangent vector at $\alpha(0)$. Then there exists a parallel vector field V along α with $V(0) = V_0$.

Proof: The expression above for $V(u)$ shows that a prospective parallel vector field V is determined by the angle $\theta(u)$. The condition that V be parallel will translate below into a complete determination of $\theta(u)$, thus constructing the desired V . The

product and chain rule give

$$\frac{d}{du}V(u) = -\sin \theta \frac{d\theta}{du}E_1 + \cos \theta \frac{d}{du}E_1 + \cos \theta \frac{d\theta}{du}E_2 + \sin \theta \frac{d}{du}E_2.$$

Using our previous calculations of the derivatives of E_1 and E_2 along α , we obtain

$$\frac{d}{du}V(u) = -\sin \theta \left[\sin v_0 + \frac{d\theta}{du} \right] E_1 + \cos \theta \left[\sin v_0 + \frac{d\theta}{du} \right] E_2 - \cos \theta \cos v_0 U.$$

Because a parallel V cannot have E_1 or E_2 components, and since $\sin \theta$ and $\cos \theta$ cannot be zero simultaneously, we must have $d\theta/du = -\sin v_0$ or

$$\begin{aligned} \theta(u) &= \theta(0) - \int \sin v_0 \, du \\ &= \theta(0) - u \sin v_0. \end{aligned}$$

This formula then defines θ and, hence, the parallel vector field V . □

Definition-Proposition. *The angle of rotation as u varies from 0 to 2π is called the holonomy along α . By the proof of the Theorem above, the holonomy along α is given by*

$$-2\pi \sin v_0.$$

Remark. Of course, all of this may be done in complete generality. Standard (and very good) references on Differential Geometry are [O’N], [Spi] and [DoC]; general results on parallelism and the *covariant derivative* may be found there.

The calculation of holonomy above says that parallel tangent vectors rotate by $-2\pi \sin v_0$ as they move completely around a latitude circle. Of course, as the

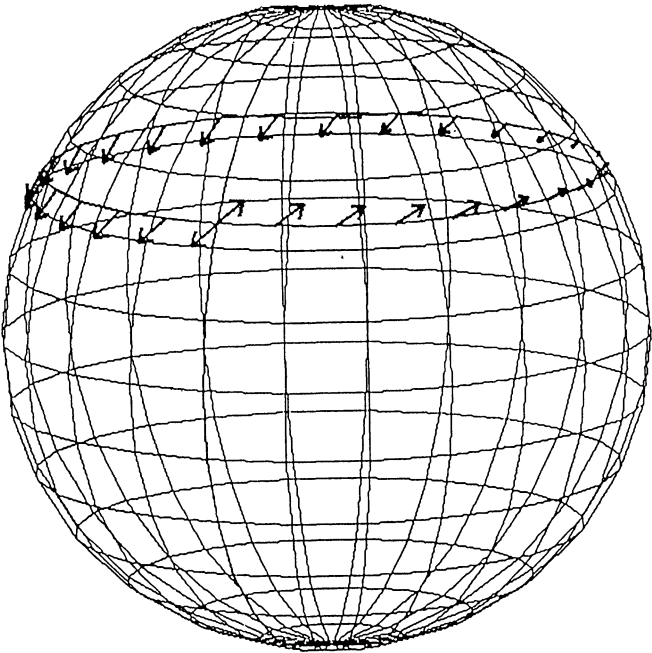


Figure 4. A Parallel Vector Field on the Sphere

terminology ‘parallel’ signifies, 2-dimensional residents of the sphere see the vectors as parallel—so, from their viewpoint, not rotating at all. This may seem contradictory since the angle between $V(u)$ and $E_1(u)$ is changing with u , but it must be remembered that the vector field E_1 along α is *not parallel*, so any angle change may be attributed to the direction change of E_1 . In fact, the product rule guarantees that two *parallel* vector fields along a curve maintain the same angle between their constituent vectors.

Exercise. What happens at the Equator and why is the Equator special among the circles of latitude?

§4. THE FOUCAULT PENDULUM. In order to analyze the Foucault pendulum from the viewpoint of geometry, assume the Earth to be non-rotating and the pendulum to be situated at latitude v_0 . Instead of the Earth rotating to move the pendulum, we move the pendulum once around the latitude circle in 24 hours at constant speed on this stationary Earth. This is clearly equivalent to the standard situation. The long cable of the pendulum and the slow progression around the latitude circle have two consequences (which are the usual physics arguments).

First, the long cable provides a relatively small swing for the pendulum which is then approximately flat. Hence, we may consider each swing as a tangent vector to the sphere. By orienting these vectors consistently, we obtain a vector field of *pendulum swing plane directions* V . At each moment of time t there is such a swing direction vector $V(t)$ and all these vectors may be placed along the latitude circle $\alpha(u)$ by associating a given moment of time t with the unique point describing the pendulum’s movement along $\alpha(u)$. Hence we write $V(u)$ for the swing plane vector field.

Secondly, because we move around the latitude circle slowly, the consequent centripetal force on the pendulum is negligible ($\approx 1/290$) compared with the downward force mg . That says that the only force F felt by the pendulum is in the normal direction U . Thus, the vertical swing plane of the pendulum experiences no tangential force and so appears unchanging to a 2-dimensional resident of the sphere. That is, *projected to the tangent plane* TS^2 ,

$$\text{proj}_{TS^2} \frac{dV(u)}{du} = 0,$$

where the covariant derivative again reduces to the ordinary derivative due to our special parametrization. By our earlier discussion, we then have

Theorem. *The vector field V associated to the Foucault pendulum is parallel along a latitude circle.*

Of course, as we transport the Foucault pendulum once around the latitude circle α , *holonomy* rotates the parallel vector field V by $-2\pi \sin v_0$ radians. In particular, the angular speed of this vector rotation is then $\omega = (2\pi \sin v_0 \text{ rads} / 24 \text{ hours})$. The equivalence of our geometric situation with the physical one then gives

Theorem. *The period of the Foucault pendulum’s precession is*

$$\frac{2\pi \text{ rads}}{\omega} = \frac{24}{\sin v_0} \text{ hours}.$$

Of course, this is precisely the period obtained in physics. Here, however, the precession of the swing-plane of the Foucault pendulum results from the holonomy along α induced by the curvature of the Earth. Further, since we view the whole pendulum apparatus as stationary relative to the Earth, what can explain the observed precession of the swing-plane? As Foucault argued, we must have

Corollary. *The Earth rotates along its latitude circles.*

Exercise. Suppose a Foucault pendulum is transported around a latitude circle on a torus. (You should still assume the only force is normal to the torus.) Compute the holonomy and explain whether this experiment alone can tell you whether we live on a sphere or torus.

Remark. While we have treated the pendulum because of its relative simplicity, a similar type of analysis can be made for one of the most useful of optimal control devices, the *gyroscope*. Indeed, in 1852 Foucault built a very refined gyroscope whose precession also demonstrated the Earth's revolution. Foucault, in fact, coined the term gyroscope from the Greek *gyros* meaning 'circle' and *skopein* meaning 'to view' because his gyroscope allowed him to see the rotation of the Earth. For more on gyroscopes see [Sca] for example.

In its own simple way, this mathematical analysis of the Foucault pendulum epitomizes the physics of the 20th century—a physics which takes a decidedly geometric view of Nature.

ACKNOWLEDGMENTS. I would like to thank Jan Oprea, Allen Broughton and John Walsh for their valuable comments and suggestions in connection with this paper. Also, the pictures of the sphere with tangent vectors and a parallel vector field were made by the software *Fields and Operators* (Lascaux Graphics). Finally, the picture of the Cleveland Museum of Natural History's Foucault pendulum was provided by Clyde Simpson of that institution.

REFERENCES

- [Ang] N. Angier, An Evolving Celebrity: Taking the Masses Beyond Dinosaurs, Interview with Stephen Jay Gould, *New York Times*, Feb. 11, 1993, p. C-1.
- [Arn] V. I. Arnold, *Mathematical Methods of Classical Mechanics 2nd Edition*, Grad. Texts in Math. 60, Springer-Verlag 1989.
- [DoC] M. Do Carmo, *Riemannian Geometry*, Birkhäuser 1992.
- [En1] M. Enos, On the Dynamics and Control of Cats, Satellites and Gymnasts: Part I, *Siam News* vol. 25 no. 5, Sept. 1992, p. 28.
- [En2] M. Enos, On the Dynamics and Control of Cats, Satellites and Gymnasts: Part II, *Siam News* vol. 25 no. 6, Nov. 1992. p. 12.
- [Mar] J. Marsden, *Lectures on Mechanics*, London Math. Soc. Lecture Note Series 174.
- [O'N] B. O'Neill, *Elementary Differential Geometry*, Academic Press 1966.
- [Opr] J. Oprea, *Differential Geometry: A Short Course*, manuscript 1993.
- [Sca] J. Scarborough, *The Gyroscope: Theory and Applications*, Interscience Publishers 1958.
- [W-S] F. Wilczek and A. Shapere (ed.), *Geometric Phases in Physics*, World Scientific, Singapore 1989.
- [Spi] Spivak, *A Comprehensive Introduction to Differential Geometry*, vol. 1–5, Publish or Perish 1971–1975.
- [Sym] K. Symon, *Mechanics 3rd Edition*, Addison-Wesley 1971.

*Department of Mathematics
Cleveland State University
Cleveland, OH 44115
oprea@csvax.csuohio.edu*

Areas of Polygons Inscribed in a Circle

David P. Robbins

1. INTRODUCTION. Since a triangle is determined by the lengths, a , b , c of its three sides, the area K of the triangle is determined by these three lengths. The well-known formula

$$K = \sqrt{s(s-a)(s-b)(s-c)}, \quad (1.1)$$

where s is the semiperimeter $(a+b+c)/2$, makes this dependence explicit. (This formula is usually ascribed to Heron of Alexandria, c. 60 BC, although some attribute it to Archimedes.)

When I was in about 7th grade I worked out Heron's formula for myself by drawing an altitude and using two instances of the Pythagorean theorem. (I was unaware of this elegant factored form above.) My fascination with the way symmetry entered the formula has stayed with me for many years.

For polygons of more than three sides, the lengths of the sides do not determine the polygon or its area. However, if we impose the condition that the polygon be convex and *cyclic*, (i.e., inscribed in a circle) then the area of the polygon is uniquely determined. Moreover, it is a symmetric function of the side lengths. The symmetry can be seen by regarding the polygon as the union of isosceles triangles each bounded by two radii and an edge of the polygon. From this point of view, we see that changing the order of the sides leaves the area unaffected. Given positive real numbers a_1, \dots, a_n , one can construct a convex n -gon with the a_i 's as the lengths of the sides provided that the largest a_j is smaller than the sum of the remaining ones. In this case it is also possible to construct a convex cyclic n -gon with the same sides and this cyclic n -gon has the largest area of all n -gons with the given side lengths. The monograph of Coxeter and Greitzer [1, pages 56–60] contains an interesting discussion which renewed my interest in the subject when I was teaching geometry at Phillips Exeter Academy almost 20 years ago.

In particular the reader will find in [1] a formula analogous to (1.1), given by Brahmagupta in the seventh century, for the area K of a cyclic quadrilateral whose four sides have lengths a , b , c , and d . It is

$$K = \sqrt{(s-a)(s-b)(s-c)(s-d)},$$

where again s is the semiperimeter $(a+b+c+d)/2$. Having read this section of [1] made me wonder what the formulas would be for polygons of more sides and I have worked sporadically on the problem since then.

In this article I will present formulas, analogous to those of Heron and Brahmagupta, for the areas of the cyclic pentagon and cyclic hexagon. For a more detailed exposition see [2].

It may seem surprising that so long a time has elapsed between the discovery of the formula for the area of the cyclic quadrilateral and the one for the cyclic pentagon. We shall see that the calculations leading to the discovery of the

pentagon formula are so complex that it would have been quite difficult to carry them out without the aid of a computer. In fact after some study of the problem I thought it likely that, even if I were to discover the formula, its complexity would make it of little interest to write down. However it is possible to write the formulas for the areas of the cyclic pentagon and the cyclic hexagon in a compact form which is related to the formula for the discriminant of a cubic polynomial in one variable.

A number of colleagues have made helpful suggestions, some of which have been incorporated in the exposition below. I would like particularly to acknowledge the contributions of Bradley Brock, Russell Kulsrud, David Lieberman, James Maiorana and Lee Neuwirth.

2. AN ALGEBRAIC FORMULATION OF THE PROBLEM. The convexity condition for polygons is algebraically unnatural. It is simpler to consider the following slightly generalized problem. Given positive real numbers a_1, \dots, a_n , find the areas of all n -gons whose side lengths are a_1, \dots, a_n and whose vertices lie on a circle. Here an n -gon is a sequence of n points P_1, \dots, P_n in a plane. Its n side lengths are the distances $P_1P_2, \dots, P_{n-1}P_n, P_nP_1$. Note that these polygons need not be convex and may intersect themselves. Let us define the area of a planar polygon whose vertices are $P_1 = (x_1, y_1), \dots, P_n = (x_n, y_n)$ to be

$$\frac{1}{2} \left(\begin{vmatrix} x_1 & y_1 \\ x_2 & y_2 \end{vmatrix} + \begin{vmatrix} x_2 & y_2 \\ x_3 & y_3 \end{vmatrix} + \dots + \begin{vmatrix} x_n & y_n \\ x_1 & y_1 \end{vmatrix} \right). \quad (2.1)$$

Defined this way the area is the sum of the areas of the components into which the polygon divides the plane, with each component weighted by the winding number of the polygon about a point in the component. The area can be negative and its sign changes when the polygon is traversed backwards. However the new formulas, like those of Heron and Brahmagupta, involve only the square of the area.

Heron's formula can be restated

$$16K^2 = 2a^2b^2 + 2a^2c^2 + 2b^2c^2 - a^4 - b^4 - c^4 \quad (2.2)$$

so that $16K^2$ is equal to a polynomial with integer coefficients in the squares of the sides of the triangle.

Brahmagupta's formula is

$$16K^2 = 2a^2b^2 + \dots + 2c^2d^2 - a^4 - b^4 - c^4 - d^4 + 8abcd$$

so that $16K^2$ is equal to a polynomial in the side lengths in which the exponents of each term are either all even or all odd.

If in the right-hand side of Brahmagupta's formula a single side length is replaced by its negative, one obtains the equation

$$16K^2 = 2a^2b^2 + \dots + 2c^2d^2 - a^4 - b^4 - c^4 - d^4 - 8abcd.$$

This has a natural geometric interpretation. It gives the area K , in the sense described above, of a nonconvex cyclic quadrilateral of side lengths a, b, c, d .

3. CONJECTURES ON AREAS OF CYCLIC POLYGONS. In this section I will describe some general conjectures about cyclic polygons. I still do not know how to prove them but the understanding they provided made it possible to discover the correct formulas for the cyclic pentagon and cyclic hexagon whose fairly simple rigorous proofs are indicated below.

Notice that Heron's formula states that $16K^2$ satisfies a monic polynomial of degree 1, namely one of the form $u - \sigma$ where σ is the right-hand side of (2.2). Its (two) coefficients are symmetric polynomials with integer coefficients in the squares of the sides of the triangle.

In the next section we shall see that, for a cyclic pentagon, $16K^2$ satisfies a monic polynomial of degree 7. Each of its coefficients is a symmetric polynomial with integer coefficients in the squares of the sides of the pentagon.

More generally there are reasons (indicated below) to believe that, for a cyclic polygon of $2m + 1$ sides, $16K^2$ satisfies a monic polynomial of degree Δ_m , where the sequence

$$\Delta_1, \Delta_2, \Delta_3, \Delta_4, \Delta_5, \dots = 1, 7, 38, 187, 874, \dots$$

is defined by

$$\Delta_m = \sum_{k=0}^{m-1} (m-k) \binom{2m+1}{k} = \frac{1}{2} \left[(2m+1) \binom{2m}{m} - 2^{2m} \right],$$

and that its coefficients are symmetric polynomials with integer coefficients in the squares of the sides of the polygon.

For polygons with $2m + 2$ sides there is an analogous conjecture that $16K^2$ satisfies one of two monic polynomials each of degree Δ_m . Both of these polynomials have coefficients which are themselves symmetric polynomials with integer coefficients in the side lengths. In these symmetric polynomials every monomial consists entirely of even powers or entirely of odd powers of the side lengths. Moreover, the two monic polynomials are closely related. Either can be obtained from the other by replacing any single side by its negative. We have already observed this for cyclic quadrilaterals and we shall see that it also holds for the hexagon.

Some parts of these conjectures are easily proved. For example, let us see why *some* algebraic relation always exists. It is helpful to employ a presentation of the problem in terms of complex numbers that will also be useful in the proofs of the formulas for the pentagon and hexagon. Assume that a polygon has its vertices on a circle centered at the origin in the complex plane. Suppose that these vertices are in order v_1, \dots, v_n and that the radius of the circle is R . Also let $v_{n+1} = v_1$ and define the quotients

$$q_j = v_{j+1}/v_j, j = 1, \dots, n.$$

Then, letting a_j be the distance from v_j to v_{j+1} , we have

$$a_j^2 = |v_{j+1} - v_j|^2 = R^2(v_{j+1} - v_j)(1/v_{j+1} - 1/v_j) = R^2(2 - q_j - q_j^{-1}). \quad (3.1)$$

Using the definition (2.1) of the area of a polygon we have

$$K = (1/2) \sum_{j=1}^n \operatorname{Im}(\bar{v}_j v_{j+1}) = (1/4i) \sum_{j=1}^n R^2(v_{j+1}/v_j - v_j/v_{j+1}).$$

Hence

$$-16K^2 = R^4(v_1/v_2 - v_2/v_1 + \dots)^2 = R^4(q_1 + \dots + q_n - q_1^{-1} - \dots - q_n^{-1})^2. \quad (3.2)$$

It follows that the $n + 1$ quantities $16K^2$ and $a_j^2, j = 1, \dots, n$, are rational functions of the $n + 1$ variables R and q_1, \dots, q_n . But the q_j 's satisfy the relation

$q_1 \cdots q_n = 1$ and are therefore algebraically dependent over the rational numbers. Hence the functions $16K^2$ and a_j^2 must themselves be algebraically dependent over the rationals.

The reason for believing that the degree of the polynomial for $(2m + 1)$ -gons is Δ_m is that Δ_m appears to be the largest number of distinct areas that can occur with a given set of side lengths.

Suppose that $n = 2m + 1$. It seems that the maximum number of areas is achieved when the n side-lengths are distinct but nearly equal. Imagine a circle of variable radius and let us try to inscribe a polygon with sides of the given lengths in the circle by picking an arbitrary starting point and laying out the edges, one at a time, with the given lengths. When the radius is too large, we will not reach the starting point when we have used up all the sides. As we decrease the radius there will come a time when we return exactly to our starting point. The resulting polygon will be nearly the regular polygon with n sides. If we continue to decrease the radius, we will overshoot the starting point starting to go around the circle again. When the radius has decreased enough, we will go around the circle exactly twice, creating a star. We can continue this way finding radii requiring more trips around the circle yielding stars with sharper points. However no edge can go as much as halfway around the circle so that the maximum number of times we can go around is m . This is where the first m areas come from.

There are other solutions. These arise as follows. We have so far assumed (implicitly) that, as we lay out all the sides around the circle, we are always proceeding in the same direction. But this is not necessary. We can lay down one of the sides in the opposite direction. Then we get a solution which looks something like a $(2m - 1)$ -gon because the backwards edge almost coincides with the preceding and following edges. Here we have $2m + 1$ choices for the backwards edge and for each of these choices we can still go around the circle $m - 1$ times. In general each choice will require a different radius and yield a different area.

This explains the term $(m - 1) \binom{2m + 1}{1}$ in the formula defining Δ_m . Subsequent terms are explained by selecting more, up to $m - 1$, of the sides to go backwards.

Diagram 1 illustrates the seven cyclic pentagons with side lengths 29, 30, 31, 32, 33.

For polygons with an even number $n = 2m + 2$ of sides, it appears that the maximum number of areas is obtained if we take $2m + 1$ of the sides to be distinct and nearly equal and the last side to be very small. We can then construct Δ_m solutions with the $2m + 1$ sides in the same orientation as above and the very small side proceeding in the same direction as the majority. We can also construct another Δ_m solutions with the very small side proceeding in the opposite direction. Thus we have a total of $2\Delta_m$ solutions, in agreement with the form conjectured above for the formula for the area of cyclic $(2m + 2)$ -gons.

4. AREA OF A CYCLIC PENTAGON. Before getting to the details of the formula, I would like to present some indication of the process by which it was discovered.

I suspected from the outset, by analogy with the case of triangles and cyclic quadrilaterals that, for a cyclic pentagon, $16K^2$ would satisfy a monic polynomial whose coefficients were symmetric polynomials in the squares of the sides of the pentagon. Considerations like those in the previous section led me to believe that the polynomial probably had degree 7. One way to check the conjecture was to

Cyclic Pentagons
Side Lengths: 29, 30, 31, 32, 33

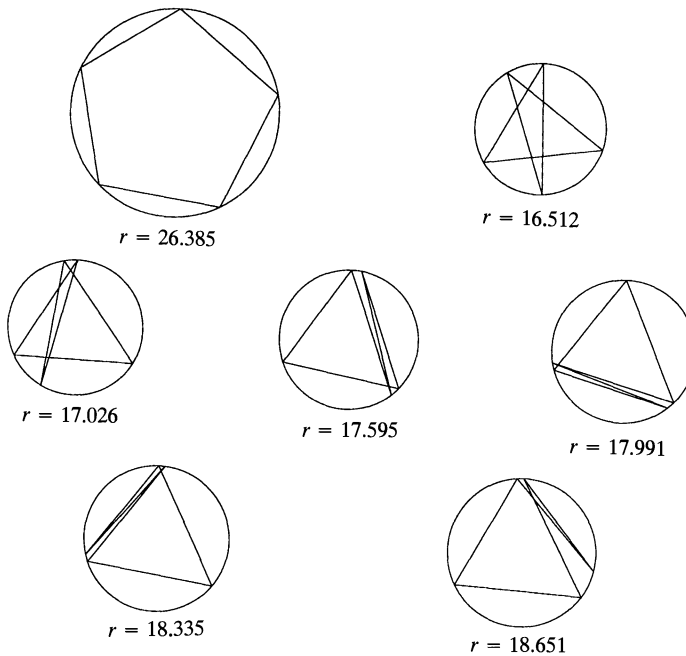


Diagram 1

choose 5 integer-valued but nearly equal sides for which all seven solutions as described above could be realized. I then computed, with (high precision) approximate arithmetic, the seven areas K of the seven pentagons described in the preceding construction and then formed the monic polynomial with the corresponding seven values of $16K^2$ as roots. If my conjecture was correct, I expected this polynomial to have integer coefficients (or near integer coefficients since the arithmetic was approximate.) It turned out that the coefficients of this polynomial were always nearly integers, as predicted.

An elaboration of this method for confirming this conjecture also leads to a computation of the correct polynomial assuming it exists. It seemed sensible to express the coefficients of the powers of $16K^2$ in terms of the elementary symmetric functions

$$\begin{aligned}\sigma_1 &= a_1^2 + \cdots + a_5^2, \\ \sigma_2 &= a_1^2 a_2^2 + \cdots + a_4^2 a_5^2 \\ &\dots\end{aligned}$$

of a_1^2, \dots, a_5^2 . Thus, taking into account the homogeneity properties of the desired polynomial, it would be of the form

$$u^7 + (c_1 \sigma_1^2 + c_2 \sigma_2) u^6 + (c_3 \sigma_1^4 + c_4 \sigma_1^2 \sigma_2 + \cdots) u^5 + \cdots$$

where $u = 16K^2$ for brevity and c_1, c_2, c_3, \dots were certain integer constants to be determined. For a given pentagon with integral sides, this polynomial could be computed exactly (rounding the near integer coefficients to the nearest integer.) Also the σ_j 's were easily found. Thus each such example gave 7 linear equations

satisfied by the c_j 's, one equation for each power of u . A simple enumeration shows that only 70 unknown c_j 's are involved in the most complicated coefficient (which is the constant term). With 70 examples (and a little luck so that the resulting systems of equations were nonsingular) it was possible to solve for the unknown coefficients.

This is how I found the formula in the first place. Many additional checks were available. One interesting check was that the computed c_j 's, which were sure to be rational from the computation method, were in fact all integers. Observe however that even though I was virtually certain that the formula was correct, I really had no proof since the formula was based on approximate arithmetic and a conjecture.

At first glance the formula looked like a random polynomial of 153 terms. However a little inspection showed that the polynomial had one very striking feature: every integer coefficient, and some were quite large, factored into *very small* primes. For example one coefficient was 2^{20} . This suggested that the polynomial had some additional structure.

By examining the polynomial carefully and manipulating it with the help of the computer program *Mathematica*, it turned out that it could be rewritten in a much more compact form as follows.

Cyclic Pentagon Area Formula. *Suppose that a pentagon inscribed in a circle has side lengths a_1, \dots, a_5 . Let $\sigma_1, \dots, \sigma_5$ be the elementary symmetric functions in the squares of the sides and let u be 16 times the square of its area. Also define t_2, t_3, t_4, t_5 by*

$$t_2 = u - 4\sigma_2 + \sigma_1^2$$

$$t_3 = 8\sigma_3 + \sigma_1 t_2$$

$$t_4 = -64\sigma_4 + t_2^2$$

$$t_5 = 128\sigma_5.$$

Then the area of a cyclic pentagon satisfies

$$ut_4^3 + t_3^2 t_4^2 - 16t_3^3 t_5 - 18ut_3 t_4 t_5 - 27u^2 t_5^2 = 0. \quad (4.1)$$

Note that, after substituting the expressions for the t_j 's in (4.1), the first term is monic of degree 7 in u and the other terms have smaller degrees in u . Hence, the formula yields a monic polynomial of degree 7 for u whose coefficients are polynomials in the squares of the lengths of the sides. The largest real root of this polynomial is 16 times the square of the area of the convex pentagon with the given side lengths.

Having discovered the Formula (4.1), I did not at first understand its significance. Some time later Bradley Brock pointed out the extremely interesting fact that the left side of (4.1) resembles the discriminant of a cubic. He was right. It turned out that it is precisely $1/(4u^2)$ times the discriminant with respect to z of the cubic polynomial

$$z^3 + 2t_3 z^2 - ut_4 z + 2u^2 t_5.$$

Why this should be the case is still a mystery. Also it should be emphasized that the quantities t_3, t_4, t_5 , which are (essentially) the coefficients of the mystery cubic, arose out of trying to make sense of the computed formula. They must have some separate significance.

The vertex quotient formulation together with this observation about the discriminant leads to a relatively simple proof of the formula which is within range of being carried out entirely by hand. (In fact a computer was used.) Recall the vertex quotients q_j and the relations (3.1) and (3.2). Let τ_1, \dots, τ_5 be the elementary symmetric functions in the q_j 's and note that $\tau_5 = q_1 \cdots q_5 = 1$. From (3.1) symmetric functions of the a_j 's are symmetric functions of the q_j 's and, from (3.2), $16K^2$ is a symmetric function in the q_j 's. Hence all the quantities t_2, t_3, t_4, t_5 and $u = 16K^2$ can be expressed in terms of τ_1, \dots, τ_4 . We easily find

$$16K^2 = -R^4(\tau_1 - \tau_4)^2.$$

Also a fairly routine but lengthy calculation, which has been omitted, yields

$$\begin{aligned} t_2 &= 4R^4(-10 + 3\tau_1 - \tau_2 - \tau_3 + 3\tau_4) \\ t_3 &= -4R^6(3\tau_1 - \tau_2 + \tau_3 - 3\tau_4)(\tau_1 - \tau_4) \\ t_4 &= 16R^8(9\tau_1 - \tau_2 + \tau_3 - 9\tau_4)(\tau_1 - \tau_2 + \tau_3 - \tau_4) \\ t_5 &= -128R^{10}(\tau_1 - \tau_2 + \tau_3 - \tau_4)^2. \end{aligned}$$

Note that each of t_3, t_4 , and t_5 factor as a product of two linear functions in the τ 's. This may be a hint for explaining the meaning of the t 's.

It is now easily verified that

$$\begin{aligned} z^3 + 2t_3z^2 - ut_4z + 2u^2t_5 &= \left[z - 16R^6(\tau_1 - \tau_4)^2 \right] \\ &\quad \times \left[z - 4R^6(\tau_1 - \tau_2 + \tau_3 - \tau_4)(\tau_1 - \tau_4) \right]^2. \end{aligned}$$

Since the cubic has a double root, we may conclude that its discriminant is 0, proving the formula.

5. AREA OF A CYCLIC HEXAGON. Similar methods can be used to find the formula for the area of cyclic hexagons. Strangely the formula can be obtained from the pentagon formula by making a slight change in the definition of the t 's.

Cyclic Hexagon Area Formula. Suppose that a hexagon inscribed in a circle has side lengths a_1, \dots, a_6 and let $\sigma_1, \dots, \sigma_5$ be the first 5 elementary symmetric functions in the squares of the sides and σ'_6 be the product of the six sides and let u be 16 times the square of its area. Also define t_2, t_3, t_4, t_5 by

$$\begin{aligned} t_2 &= u - 4\sigma_2 + \sigma_1^2 \\ t_3 &= 8\sigma_3 + \sigma_1t_2 - 16\sigma'_6 \\ t_4 &= t_2^2 - 64\sigma_4 + 64\sigma_1\sigma'_6 \\ t_5 &= 128\sigma_5 + 32t_2\sigma'_6. \end{aligned}$$

Then the area of a cyclic hexagon satisfies either

$$ut_4^3 + t_3^2t_4^2 - 16t_3^3t_5 - 18ut_3t_4t_5 - 27u^2t_5^2 = 0 \quad (5.1)$$

or the equation obtained by replacing σ'_6 by its negative.

The reader may wonder how the squared area of a given hexagon decides which of these two equations to solve. The answer is to look at the product

$$p = (1 - q_1) \cdots (1 - q_6),$$

where q_j 's are the vertex quotients. Since each q_j has absolute value 1, and the product of the q_j 's is 1, it is easily verified that p is always real. A squared area is a root of (5.1) when $p > 0$ and a root of the alternate form if $p < 0$. In particular the convex case yields a root of (5.1).

6. AREA OF A CYCLIC HEPTAGON. These methods could be used in principle to derive the (degree 38) formula for the cyclic heptagon squared area. However the computations would be of rather heroic proportions, requiring for some of the coefficients the solution of a system of linear equations with 143307 unknowns. Perhaps someone can guess an answer like the compact formulas for the pentagon and hexagon, which might then be provable with a simple argument as above.

REFERENCES

1. H. S. M. Coxeter and S. L. Greitzer, *Geometry Revisited*, The Mathematical Association of America, (1967).
2. David P. Robbins, Areas of Polygons Inscribed in a Circle, *Discrete and Computational Geometry*, 12: 223–236 (1994).

*Center for Communications Research
Thanet Road
Princeton, NJ 08540
robbins@ccr-p.ida.org*

PICTURE PUZZLE

(from the collection of Paul Halmos)



They have the same name. Well, not quite.
(see page 537.)

Curves of Constant Precession

Paul D. Scofield

1. INTRODUCTION. Given initial position and direction, the flight-path of a ship in Euclidean space is completely determined by how much it turns and how much it twists at each odometer reading. This is an intuitive interpretation of the Fundamental Theorem for Space Curves, which states that curvature κ and torsion τ , as functions of arclength s , determine a space curve uniquely up to rigid motion. This statement of the Fundamental Theorem ([14], §1–8) should be tempered with the reservations expressed by Nomizu [12] and Wong & Lai [15].

Given a parametric space curve, there are well-known formulae for the arclength, curvature, and torsion (as functions of the parameter). Given two functions of one parameter (potentially curvature and torsion parametrized by arc-length) one might like to find a parametrized space curve for which the two functions are the curvature and torsion. This activity, called “solving natural equations” ([14], §1–10), is generally achieved by solving Riccati equations like $dw/ds = -i\tau/2 - i\kappa w + i\tau w^2/2$.

Although the solution generally exists, it usually cannot be obtained explicitly. Euler [6] found explicit integral formulae for plane curves (where $\tau \equiv 0$) through direct geometric analysis. Hoppe [9] developed a general method for solving the natural equations for space curves by solving Riccati equations through a complicated sequence of integral transformations. He digressed to obtain formulae for the tangent, normal, and binormal indicatrices for general helices and essentially for curves of constant precession. Enneper [5] obtained explicit closed-form solutions for helices on revolved conic sections through direct geometric analysis.

A curve of constant precession is defined by the property that as the curve is traversed with unit speed, its centrode revolves about a fixed axis with constant angle and constant speed. In this paper we obtain an arclength-parametrized closed-form solution of the natural equations for curves of constant precession through direct geometric analysis. As part of this analysis, we obtain a new theorem for curves of constant precession analogous with Lancret’s Theorem for general helices. We provide the first rendering of a curve of constant precession. We also note for the first time that curves of constant precession lie on circular hyperboloids of one sheet and have closure conditions that are simply related to their arclength, curvature, and torsion. These are 3-type curves, except one family of closed 2-type curves (when $\omega = \sqrt{3}\mu$; see [2], [3], and [1]).

Given a closed C^3 curve in space, it is rather obvious that the curvature and torsion functions will be periodic functions of the arclength, with period equal the total arclength. This is a necessary condition but, as the circular helices (κ and τ both constant) show, not a sufficient condition that integral curves be closed. Efimov [4] and Fenchel [7] independently formulated

The Closed Curve Problem. *Find (explicit) necessary and sufficient conditions that determine when, given two periodic functions $\kappa(s)$ and $\tau(s)$ with the same period L , the integral curve is closed.*

This natural problem in elementary differential geometry remains open, despite implicit solutions by Schmeidler [13] and Hwang [10]. Fenchel warned that there may be no simple solution. Our investigation of curves of constant precession began in an effort to find closure conditions for some collection of pairs of simple periodic functions like $\kappa(s) = \omega \cos \mu s$ and $\tau(s) = \omega \sin \mu s$.

2. PLANE CURVES. Here we set out Euler's well-known integral solutions of the natural equations for plane curves ([14], p. 26). We will designate coordinates and geometric invariants of plane curves by subscript π . Identifying the angle between the tangent line to the curve and the x -axis as

$$\varphi_\pi = \int \kappa_\pi ds_\pi,$$

it follows that

$$x_\pi = \int \cos \varphi_\pi ds_\pi \quad \text{and} \quad y_\pi = \int \sin \varphi_\pi ds_\pi$$

solve natural equations of the form

$$\kappa_\pi = \kappa_\pi(s_\pi) \quad \text{and} \quad \tau_\pi \equiv 0.$$

If we change a constant of integration, we rotate or translate the curve.

Still, it is a rare curve for which both κ is a simple function and the above integrals can be evaluated in closed form with elementary functions. Among the simplest are the circle, the logarithmic spiral, the circle involute, and the epicycloid ([14], pp. 26–28). Enneper [5] showed that each of these is the projection along the axis of symmetry of a curve of constant slope (helix) on a conic surface of revolution: a circular cylinder, a cone, a paraboloid, and a sphere.

3. CURVES OF CONSTANT SLOPE (HELICES). Here we set out the integral solution of the natural equations for curves of constant slope or general helices ([14], pp. 33–35), and we set out an explicit parametrization for spherical helices, never appealing to the solution of a Riccati equation. A *curve of constant slope* or *helix* is defined by the property that the tangent makes a constant angle θ with a fixed line l . We have its natural equations by

The Theorem of Lancret [11]. *A necessary and sufficient condition that a curve be of constant slope is that the ratio of curvature to torsion be constant.*

In proving the theorem, it is observed that the constant slope and the constant ratio are related by

$$\kappa/\tau = \tan \theta, \text{ constant.}$$

Taking l as the z -axis, it is easy to observe that $dz = \cos \theta ds$. Moreover, the projection of the curve onto the xy -plane has arclength element $ds_\pi = \sin \theta ds$ and curvature $\kappa_\pi = \kappa \csc^2 \theta$ (relating the radii of a helical osculating circle and the planar osculating circle of its projection). Then using Euler's planar solution,

$$\varphi_\pi = \int \kappa_\pi ds_\pi = \csc \theta \int \kappa ds,$$

so

$$\begin{aligned} x(s) &= \sin \theta \int_0^s \cos \left[\csc \theta \int_0^{s_1} \kappa(s_2) ds_2 \right] ds_1 \\ y(s) &= \sin \theta \int_0^s \sin \left[\csc \theta \int_0^{s_1} \kappa(s_2) ds_2 \right] ds_1 \\ z(s) &= s \cos \theta. \end{aligned}$$

General helices are precisely the geodesics on general cylinders generated by lines parallel with l . A general cylinder is the rectifying developable of its helices.

We will want a parametrization for spherical helices because the tangent indicatrix of a curve of constant precession will prove to be a spherical helix. In anticipation, we will designate the coordinates and arclength of spherical helices by subscript t . Struik ([14], pp. 34–35) shows that for a helix on a sphere of radius r making an angle θ with the z -axis, the projection onto the xy -plane is an epicycloid with fixed radius $a = r \cos \theta$ and rolling radius $b = r \sin^2(\theta/2)$. Substituting these into his epicycloid parametrization (p. 27), we obtain

$$\begin{aligned} x_t(\psi) &= \frac{r}{2}(1 + \cos \theta) \cos \psi - \frac{r}{2}(1 - \cos \theta) \cos \frac{1 + \cos \theta}{1 - \cos \theta} \psi \\ y_t(\psi) &= \frac{r}{2}(1 + \cos \theta) \sin \psi - \frac{r}{2}(1 - \cos \theta) \sin \frac{1 + \cos \theta}{1 - \cos \theta} \psi \\ z_t(\psi) &= r \sin \theta \cos \frac{\cos \theta}{1 - \cos \theta} \psi, \end{aligned}$$

where

$$\begin{aligned} s_\pi &= r \sin \theta \tan \theta \cos \frac{\cos \theta}{1 - \cos \theta} \psi, \\ s_t &= r \tan \theta \cos \frac{\cos \theta}{1 - \cos \theta} \psi. \end{aligned}$$

The spherical helix has an arc of length $2r \tan \theta$ between heights $z = \pm r \sin \theta$ beyond which no tangent to the sphere makes an angle as small as θ with the z -axis. The parametric extension gives a sequence of arcs which join in cusps at their endpoints. This piecewise smooth curve is closed if and only if $\cos \theta$ is rational. All arcs of a spherical helix with $\cos \theta = 8/17$ are rendered in Figure 1.

4. CURVES OF CONSTANT PRECESSION. Here we characterize curves of constant precession. We will denote the moving orthonormal frame of tangent, normal, and binormal vectors by \mathbf{t} , \mathbf{n} , and \mathbf{b} , and we will differentiate with respect to arclength, using the *Frenet equations*

$$\begin{aligned} \mathbf{t}' &= \kappa \mathbf{n} \\ \mathbf{n}' &= -\kappa \mathbf{t} + \tau \mathbf{b} \\ \mathbf{b}' &= -\tau \mathbf{n} \end{aligned}$$

([8], [14], §1-6). Let $\mathbf{C} = \tau \mathbf{t} + \kappa \mathbf{b}$ denote the *centrode*, the Frenet frame's axis of instantaneous rotation ([14], §1-6, Exercise 18, and [7]). Fix arbitrary constants $\omega > 0$, μ , and $\alpha = \sqrt{\omega^2 + \mu^2}$. Set $\mathbf{A} = \mathbf{C} \pm \mu \mathbf{n}$ and fix the line l parallel with $\mathbf{A}(0)$. We use $\angle(\bullet, \bullet)$ to denote the angle between two vectors.

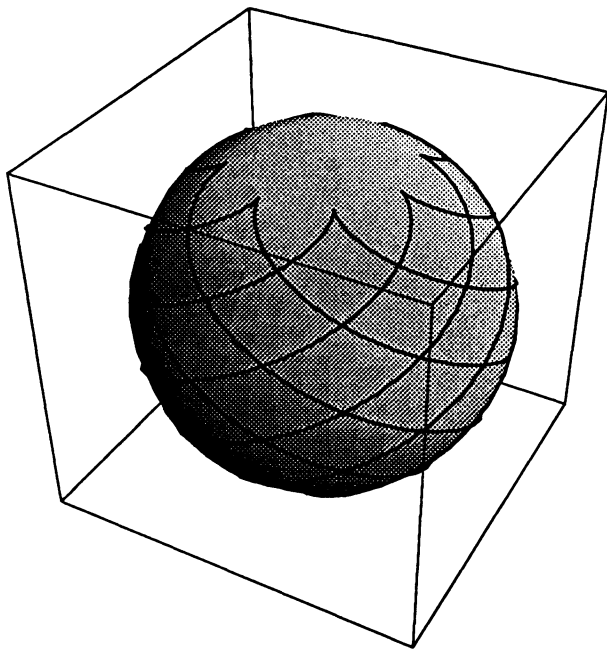


Figure 1. Sixteen arcs of a helix on the unit sphere with $\cos \theta = 8/17$. They form the tangent indicatrix of the curve in Figure 2. (Visualization assisted by The Geometry Center at the University of Minnesota.)

Lemma. *The following are equivalent:*

- (i) $|\mathbf{C}| = \omega$
- (ii) $\angle(\mathbf{C}, \mathbf{A}) = \cos^{-1} \frac{\omega}{\alpha}$
- (iii) $|\mathbf{n}'| = \omega$
- (iv) $\angle(\mathbf{n}, \mathbf{A}) = \cos^{-1} \frac{\mu}{\alpha}$
- (v) $|\mathbf{A}| = \alpha$.

Proof: Since $|\mathbf{C}|^2 = \kappa^2 + \tau^2 = |\mathbf{n}'|^2$ and $|\mathbf{A}|^2 = \kappa^2 + \tau^2 + \mu^2$, it is clear that (i), (iii) and (v) are equivalent. Interpreting (ii) as

$$\kappa^2 + \tau^2 = \mathbf{C} \cdot \mathbf{A} = \frac{\omega}{\alpha} \sqrt{\kappa^2 + \tau^2} \sqrt{\kappa^2 + \tau^2 + \mu^2}$$

implies that (i) is equivalent to (ii), and interpreting (iv) as

$$\mu = \mathbf{n} \cdot \mathbf{A} = \frac{\mu}{\alpha} |\mathbf{n}| |\mathbf{A}|$$

implies that (iv) is equivalent to (v). Q.E.D.

Lemma. *Given any of (i)–(v), the following are equivalent:*

- (vi) $|\mathbf{C}'| = |\omega\mu|$
- (vii) \mathbf{A} is parallel with l .

Proof: Since $\mathbf{A}' = \mathbf{C}' \pm \mu \mathbf{n}'$,

$$\mathbf{A}' = 0 \Leftrightarrow \mathbf{C}' = \mp \mu \mathbf{n}' \Leftrightarrow |\mathbf{C}'| = |\mu| |\mathbf{n}'|.$$

Thus, it follows from (iii) and (v) that (vi) and (vii) are equivalent. Q.E.D.

A curve of constant precession is defined (somewhat redundantly) by the property that, as it is traversed with unit speed, its centrode revolves about a fixed line l in space (the *axis*) with constant angle and constant speed. As a consequence, its Frenet frame precesses about l , while its principal normal revolves about l with constant complementary angle and constant speed. We have its natural equations by the following analogy with Lancret's Theorem.

Theorem 1. *A necessary and sufficient condition that a curve be of constant precession is that $\kappa(s) = \omega \sin \mu s$ and $\tau(s) = \omega \cos \mu s$, up to reflection or phase shift of arclength, for constants ω and μ .*

Proof: Conditions (v) and (vii) are true if and only if $\mathbf{A}' = 0$, but

$$\mathbf{A}' = (\tau' - \mu \kappa) \mathbf{t} + (\kappa' + \mu \tau) \mathbf{b}$$

and uniqueness of solutions of pairs of linear equations imply that $\mathbf{A}' = 0$ if and only if $\kappa(s) = \omega \sin \mu s$ and $\tau(s) = \omega \cos \mu s$ (up to reflection or phase shift). Q.E.D.

5. SOLVING THE NATURAL EQUATIONS. Here, without solving a Riccati equation but using results from Sections 3 and 4, we obtain an arclength parametrization for curves of constant precession. Condition (iv) of the lemmata in Section 4 implies, since $\mathbf{t}' = \kappa \mathbf{n}$, that \mathbf{t} is a curve of constant slope (hence a helix on the unit sphere). We take $\kappa = \pm \omega \sin \mu s$ and continue to designate the tangent indicatrix by subscript t . Arclength along the curve and along its tangent indicatrix are related by

$$\frac{ds_t}{ds} = \kappa = \pm \omega \sin \mu s,$$

so

$$s_t = \mp \frac{\omega}{\mu} \cos \mu s + C.$$

Taking the lower signs, $C = 0$, and $\alpha = |\mathbf{A}| = \sqrt{\omega^2 + \mu^2}$, while substituting $r = 1$ and $\cos \theta = \mu/\alpha$ into the formula for s_t in Section 3, we obtain

$$s_t = \frac{\omega}{\mu} \cos \frac{\mu}{\alpha - \mu} \psi$$

hence

$$s = \frac{1}{\alpha - \mu} \psi,$$

giving a remarkably simple reparametrization

$$\begin{aligned} x'(s) &= x_t(s) = \frac{\alpha + \mu}{2\alpha} \cos(\alpha - \mu)s - \frac{\alpha - \mu}{2\alpha} \cos(\alpha + \mu)s \\ y'(s) &= y_t(s) = \frac{\alpha + \mu}{2\alpha} \sin(\alpha - \mu)s - \frac{\alpha - \mu}{2\alpha} \sin(\alpha + \mu)s \\ z'(s) &= z_t(s) = \frac{\omega}{\alpha} \cos \mu s. \end{aligned}$$

Theorem 2. An arclength parametrization of a curve of constant precession with natural equations $\kappa(s) = -\omega \sin \mu s$ and $\tau(s) = \omega \cos \mu s$ is given by

$$\begin{aligned} x(s) &= \frac{\alpha + \mu}{2\alpha} \frac{\sin(\alpha - \mu)s}{\alpha - \mu} - \frac{\alpha - \mu}{2\alpha} \frac{\sin(\alpha + \mu)s}{\alpha + \mu} \\ y(s) &= -\frac{\alpha + \mu}{2\alpha} \frac{\cos(\alpha - \mu)s}{\alpha - \mu} + \frac{\alpha - \mu}{2\alpha} \frac{\cos(\alpha + \mu)s}{\alpha + \mu} \\ z(s) &= \frac{\omega}{\mu\alpha} \sin \mu s \end{aligned}$$

where ω , μ , and $\alpha = \sqrt{\omega^2 + \mu^2}$ are constant. Moreover, the curve lies on the circular hyperboloid of one sheet

$$x^2 + y^2 - \frac{\mu^2}{\omega^2} z^2 = \frac{4\mu^2}{\omega^4}.$$

The curve is closed if and only if μ/α is rational.

A curve of constant precession is rendered in Figure 2. The tangent indicatrix, a spherical helix, has cusps where $\kappa(s) = -\omega \sin \mu s = 0$.

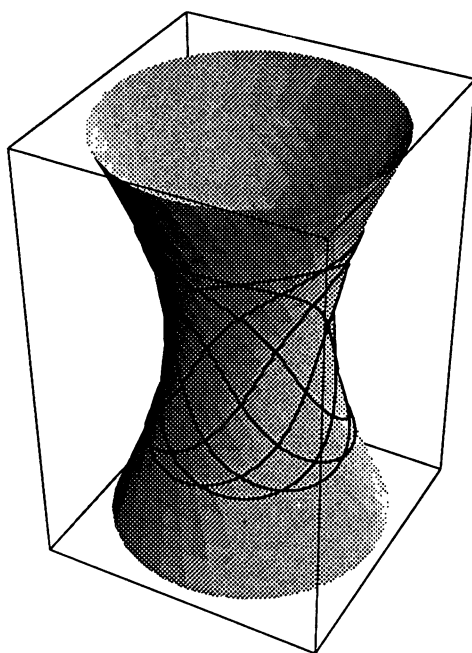


Figure 2. A curve of constant precession with $\omega = 15$ and $\mu = 8$, shown on its circular hyperboloid. It is an integral curve of the indicatrix in Figure 1.

REFERENCES

1. B.-Y. Chen, Some open problems and conjectures on submanifolds of finite type, *Soochow J. Math.* 17 (1991), 169–188.
2. B.-Y. Chen, J. Deprez, F. Dillen, L. Verstraelen, and L. Vrancken, Curves of finite type, in *Geometry and Topology of Submanifolds, II*, World Scientific, Singapore, 1990, pp. 76–110.

3. J. Deprez, F. Dillen, and L. Vrancken, Finite type curves on quadrics, *Chinese J. Math* 18 (1990), 95–121.
4. N. V. Efimov, Nekotorye zadachi iz teorii prostranstvennykh krivyykh, *Usp. Mat. Nauk* 2 (1947), 193–194.
5. A. Enneper, Zur Theorie der Curven doppelter Krümmung, *Math. Ann.* 19 (1882), 72–83.
6. L. Euler, De constructione aequationum ope motus tractorii aliisque ad methodum tangentium inversam pertinentibus, *Comm. Acad. Sti. Petrop.* 8 (1736), 66–85. Reprinted in *Leonhardi Euleri Opera Omnia, Ser. I*, Vol. 22, Basel, 1926, pp. 83–107.
7. W. Fenchel, The differential geometry of closed space curves, *Bull. Amer. Math. Soc.* 57 (1951), 44–54.
8. F. Frenet, Sur les courbes à double courbure, *Jour. de Math.* 17 (1852), 437–447. Extrait d'une Thèse (Toulouse, 1847).
9. R. Hoppe, Ueber die Darstellung der Curven durch Krümmung und Torsion, *J. Reine Angew. Math.* 60 (1862), 182–187.
10. C.-C. Hwang, A differential-geometric criterion for a space curve to be closed, *Proc. Amer. Math. Soc.* 83 (1981), 357–361.
11. M. A. Lancret, Mémoire sur les courbes à double courbure, *Mém. des sav. étrangers* 1 (1806), 416–454.
12. K. Nomizu, On Frenet equations for curves of class C^∞ , *Tôhoku Math. J.* (2) 11 (1959), 106–112.
13. W. Schmeidler, Notwendige und hinreichende Bedingungen dafür, daß eine Raumkurve geschlossen ist, *Arch. Math.* 7 (1956), 384–385.
14. D. J. Struik, *Lectures on Classical Differential Geometry*, Dover, New York, 1988. Reprint of second edition (Reading, 1961).
15. Y.-C. Wong and H.-F. Lai, A critical examination of the theory of curves in three dimensional geometry, *Tôhoku Math. J.* (2) 19 (1967), 1–31.

Department of Mathematics
Washington and Lee University
Lexington, VA 24450
scofield@wlu.edu

Don't talk to me of your Archimedes' lever. He was an absentminded person with a mathematical imagination. Mathematics commands all my respect, but I have no use for engines. Give me the right word and the right accent and I will move the world.

—*Joseph Conrad*

Preface to *A Personal Record*. Garden City NY: Doubleday, Doran and Co. Inc., 1929, p. xiii.

**Answer to Picture Puzzle
 (p. 530)**

Carl Ludwig Siegel and Grahame Segal.

NOTES

Edited by: John Duncan

Matrix Expansion by Orthogonal Kronecker Products

Jeffery C. Allen

The singular-value decomposition (SVD) provides an expansion of a real $M \times N$ matrix A by orthogonal outer products [3, page 144]:

$$A = \sum_{k=1}^K s_k \mathbf{u}_k \mathbf{v}_k^T.$$

The singular values are ordered $s_1 \geq s_2 \geq \cdots \geq s_K \geq 0$, where $K = \min\{M, N\}$. The \mathbf{u}_k 's and the \mathbf{v}_k 's are the orthonormal left and right singular vectors:

$$\mathbf{u}_k^T \mathbf{u}_{k'} = \mathbf{v}_k^T \mathbf{v}_{k'} = \begin{cases} 1 & k = k' \\ 0 & k \neq k' \end{cases},$$

where \mathbf{u}^T denotes the vector transpose. This note demonstrates that by a simple *rescanning* of the matrix A , the SVD can also produce a variety of expansions in terms of orthogonal Kronecker products.

NOTATION. Vectors are denoted by a bold lowercase letter. The “vec” of a matrix is the vector that results by stacking the columns. For example,

$$\text{vec}([\mathbf{x}_1, \mathbf{x}_2]) = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}.$$

$\text{Mat}(M, N)$ denotes the linear space of real $M \times N$ matrices. $\text{BlkMat}(M_1, N_1; M_2, N_2)$ denotes the linear space of real $M_1 \times N_1$ block matrices with $M_2 \times N_2$ blocks. The inner product of two matrices is given by $\langle U, V \rangle = \text{vec}(U)^T \text{vec}(V)$. The Kronecker product of two matrices is $U \otimes V = [u_{m,n} V]$ [3, page 243]. For example,

$$\begin{bmatrix} u_{1,1} & u_{1,2} \\ u_{2,1} & u_{2,2} \end{bmatrix} \otimes V = \begin{bmatrix} u_{1,1}V & u_{1,2}V \\ u_{2,1}V & u_{2,2}V \end{bmatrix}.$$

Note that in $U \otimes V$, every element of U multiplies every element of V . Thus, the Kronecker product contains the same products as the outer product $\text{vec}(V)\text{vec}(U)^T$. The map taking Kronecker products to outer products is the rescanning function.

THE RESCANNING FUNCTION. Let M_1, M_2, N_1 , and N_2 be positive integers. Use these to define the mapping **block**: $\text{Mat}(M_2 N_2, M_1 N_1) \rightarrow \text{BlkMat}(M_1, M_2; N_1, N_2)$

$$\mathbf{block}([\mathbf{a}_1, \dots, \mathbf{a}_{M_1 N_1}]) = \begin{bmatrix} A_1 & A_{M_1+1} & \cdots & A_{(N_1-1)M_1+1} \\ A_2 & A_{M_1+2} & \cdots & A_{(N_1-1)M_1+2} \\ \vdots & \vdots & \ddots & \vdots \\ A_{M_1} & A_{2M_1} & \cdots & A_{N_1 M_1} \end{bmatrix},$$

where each block A_k is a real $M_2 \times N_2$ matrix determined from \mathbf{a}_k by $\text{vec}(A_k) = \mathbf{a}_k$. Thus, **block** merely *rescans* a matrix of size $M_2 N_2 \times M_1 N_1$ into an $M_1 \times N_1$ block matrix with $M_2 \times N_2$ blocks. It is straight-forward to verify the following:

(B-1) **block** is a linear mapping which is one-to-one and onto.

(B-2) **block** preserves the matrix inner product.

(B-3) **block** maps outer products onto Kronecker products.

To illustrate the last claim, let $U \in \text{Mat}(M_1, N_1)$ and $V \in \text{Mat}(M_2, N_2)$. Then $A = U \otimes V$ is a $M_1 \times N_1$ block matrix with $M_2 \times N_2$ blocks $A_{m+(n-1)M_1} = u_{m,n} V$ for $m = 1, \dots, M_1$ and $n = 1, \dots, N_1$. In matrix form, this can be written

$$\mathbf{block}(\text{vec}(V)\text{vec}(U)^T) = U \otimes V.$$

Since **block** is an isometry mapping outer products to Kronecker products, it should come as no surprise that **block** lifts the SVD expansion to the Kronecker expansion.

THE KRONECKER EXPANSION. Let $A \in \text{Mat}(M, N)$ where $M = M_1 M_2$ and $N = N_1 N_2$. Let $K = \min\{M_1 N_1, M_2 N_2\}$. Then there are matrices $U_1, \dots, U_K \in \text{Mat}(M_1, N_1)$, matrices $V_1, \dots, V_K \in \text{Mat}(M_2, N_2)$, and numbers $s_1 \geq \dots \geq s_K \geq 0$ such that

$$A = \sum_{k=1}^K s_k U_k \otimes V_k$$

where the following orthogonality conditions hold:

$$\langle U_k, U_{k'} \rangle = \langle V_k, V_{k'} \rangle = \begin{cases} 1 & k = k' \\ 0 & k \neq k' \end{cases}.$$

Proof: By (B-1), **block** is invertible. Set $\mathbf{A} = \mathbf{block}^{-1}(A)$. Write the SVD of matrix \mathbf{A} in the form

$$\mathbf{A} = \sum_{k=1}^K s_k \mathbf{u}_k \mathbf{u}_k^T.$$

Let $U_k \in \text{Mat}(M_1, N_1)$ be determined from $\text{vec}(U_k) = \mathbf{u}_k$. Likewise, let $V_k \in \text{Mat}(M_2, N_2)$ be determined from $\text{vec}(V_k) = \mathbf{v}_k$. Then by (B-1) and (B-3), the

Kronecker expansion is obtained

$$A = \mathbf{block}(A) = \sum_{k=1}^K s_k \mathbf{block}(\mathbf{v}_k \mathbf{u}_k^T) = \sum_{k=1}^K s_k U_k \otimes V_k.$$

By (B-2), it is straight-forward to verify the orthogonality of the U_k 's and the V_k 's.

Remarks. Since the SVD also determines optimal reduced rank approximations, best approximations using a fixed number of Kronecker products can be obtained from this Kronecker expansion. The motivation for this Kronecker expansion or “block” SVD arose in an image-processing application. Illuminating discussions of image processing and applications of the SVD, block matrix computations, and Kronecker products are found in Gonzalez and Wintz [6] or Jain [5]. An excellent treatment of the Kronecker product is found in Horn and Johnson [3]. Further generalizations of the Kronecker product and signal-processing applications are found in [4], [7], [2], [1].

REFERENCES

1. Andrews, H. C. and J. Kane, Kronecker Matrices, Computer Implementation, and Generalized Spectra, *Journal of the Association for Computing Machinery*, Volume 17, Number 2, (1970) 260–268.
2. Henderson, Harold V. and S. R. Searle, The Vec-Permutation Matrix, the Vec Operator and Kronecker Products: A Review, *Linear and Multilinear Algebra*, Volume 9 (1981), 271–288.
3. Horn, Roger A. and Charles R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge (1991).
4. Hyland, David C. and Emmanuel G. Collins, Jr., Block Kronecker Products and Block Norm Matrices in Large-Scale Systems Analysis, *SIAM Journal of Matrix Analysis and Applications*, Volume 10, Number 1 (1989) 18–29.
5. Jain, Anil K. *Fundamentals of Digital Image Processing*, Prentice Hall (1989).
6. Gonzalez, Rafael C. and Paul Wintz, *Digital Image Processing*, Addison-Wesley Publishing Company (1977).
7. Regalia, Phillip A. and Sanjuit K. Mitra, Kronecker Products, Unitary Matrices and Signal Processing Applications, *SIAM Review*, Volume 31, Number 4 (1989) 586–613.

NCCOSC RDTE DIV 574
53560 Hull Street
San Diego, CA 92152-5001
allen@nosc.mil

Injective Polynomial Maps Are Automorphisms

Walter Rudin

This article presents a simple elementary proof of the following result.

Theorem A. *If $F: \mathbb{C}^n \rightarrow \mathbb{C}^n$ is a polynomial map which is one-to-one, then*

- (a) $F(\mathbb{C}^n) = \mathbb{C}^n$, and
- (b) $F^{-1}: \mathbb{C}^n \rightarrow \mathbb{C}^n$ is also a polynomial map.

Here n is a positive integer, and \mathbb{C}^n is the set of all $z = (z_1, \dots, z_n)$, each z_i lying in the complex field \mathbb{C} . In general, the notation $\Phi: X \rightarrow Y$ indicates that Φ is a map whose domain is X and whose range lies in Y . To say that F is a *polynomial* map means that $F = (f_1, \dots, f_n)$ and each component f_i of F is a polynomial, mapping \mathbb{C}^n into \mathbb{C} .

Theorem A may be regarded as a small step toward a confirmation of the so-called Jacobian conjecture, which claims that if $F: \mathbb{C}^n \rightarrow \mathbb{C}^n$ is a polynomial map whose Jacobian is a non-zero constant, then F is a polynomial automorphism of \mathbb{C}^n , i.e., F is one-to-one and satisfies (a) and (b). This dates back to 1939 [5] but is still unproved (in June 1994), even for $n = 2$. Its history, many references, and some partial results, can be found in [2].

Theorem A shows that the Jacobian conjecture would be proved if one could show, for polynomial maps $F: \mathbb{C}^n \rightarrow \mathbb{C}^n$, that “locally one-to-one” implies “globally one-to-one.” This formulation of the problem points to an interesting difference between \mathbb{C}^n and \mathbb{R}^n : Serguey Pinchuk [8] has (surprisingly!) constructed a polynomial map $F: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ whose Jacobian has no zero in \mathbb{R}^2 but which is not one-to-one. The difference is, of course, that on \mathbb{R}^n there are nonconstant polynomials without zeros, whereas this cannot happen on \mathbb{C}^n .

Theorem A is not new. In [7] Don Newman proved (a) with \mathbb{R}^2 in place of \mathbb{C}^n . In [3] this was extended to \mathbb{R}^n , for arbitrary n , with the aid of a good dose of homology theory; that paper also contains a brief sketch of the analogous result for maps from k^n to k^n , for arbitrary algebraically closed fields k . Ax [1; Th. 2] extended this to morphisms of algebraic varieties, using nonprincipal ultraproducts of fields. Theorem (2.1) on p. 294 of [2] lists eight (mostly algebraic) conditions on polynomial maps F that are equivalent; Theorem A is one of those equivalences: F is one-to-one if and only if F is an automorphism.

I believe that the proof given here is much simpler than any of the above. (Proof: I have no trouble understanding it.) It uses two facts from complex analysis:

Fact 1. If (i) $u, v: \mathbb{C}^n \rightarrow \mathbb{C}$ are polynomials with no common factor of positive degree,

(ii) Ω is an open subset of \mathbb{C}^n , and

(iii) $v(p_0) = 0$ at some point p_0 in Ω ,

then Ω contains points p at which $v(p) = 0$ but $u(p) \neq 0$.

This must be prehistoric. A proof can be found on pp. 14, 15 of [11]. Note that it fails on \mathbb{R}^n .

Example: $u(x, y) = x^2 + y^2$, $v(x, y) = x^2 + (y - x)^2$.

Fact 2. If F satisfies the hypothesis of Theorem A, then the Jacobian of F is $\neq 0$ at every point of \mathbb{C}^n .

This is in fact true for holomorphic maps from open sets in \mathbb{C}^n into \mathbb{C}^n that are locally one-to-one, and it used to be a fairly difficult theorem (see, for instance, [6; pp. 86–88]) until Jean-Pierre Rosay published a truly simple proof [9].

Combined with the inverse function theorem (Th. 9.24 in [10]), Fact 2 implies what will actually be used, namely:

The range $F(\mathbb{C}^n)$ of F is an open subset of \mathbb{C}^n .

(Remark: That $F(\mathbb{C}^n)$ is open is also an immediate consequence of Brouwer's "Invariance of Domain" theorem, concerning continuous one-to-one maps from \mathbb{R}^N into \mathbb{R}^N [4; p. 95] but that theorem is much more difficult than the route via Fact 2.)

We now start the proof.

Let f_1, \dots, f_n be the components of F , and let k be the subfield of \mathbb{C} generated by the coefficients of the polynomials f_i . Since k is countable, there are only countably many polynomials with coefficients in k . The union of their zero-sets (ignoring the zero-polynomial) is thus a countable union of closed sets without interior, hence cannot cover the complete metric space \mathbb{C}^n . It follows that there is a point ξ in \mathbb{C}^n , fixed from now on, with the following property:

(*) If $f: \mathbb{C}^n \rightarrow \mathbb{C}$ is a polynomial with coefficients in k , and $f(\xi) = 0$, then $f(z) = 0$ for every z in \mathbb{C}^n .

Put $\eta = F(\xi)$.

Claim. *The extension fields*

$$k(\eta) = k(\eta_1, \dots, \eta_n)$$

and

$$k(\eta, \xi) = k(\eta_1, \dots, \eta_n, \xi_1, \dots, \xi_n)$$

are equal.

Here $k(\eta)$ is the smallest subfield of \mathbb{C} that contains k and η_1, \dots, η_n , and similarly for $k(\eta, \xi)$.

If the claim is false, there is an isomorphism φ of $k(\eta, \xi)$ into \mathbb{C} that fixes every element of $k(\eta)$ but moves some ξ_i . (See the lemma at the end of the paper.) Put

$$\omega = (\varphi(\xi_1), \dots, \varphi(\xi_n))$$

and note that $\omega \neq \xi$.

Since $f_j(\xi) = \eta_j$ is in $k(\eta)$ and the coefficients of f_j are in k , we have, for $1 \leq j \leq n$,

$$f_j(\xi) = \varphi(f_j(\xi_1, \dots, \xi_n)) = f_j(\varphi(\xi_1), \dots, \varphi(\xi_n)) = f_j(\omega).$$

Hence $F(\xi) = F(\omega)$, which contradicts the assumption that F is one-to-one. This proves the claim.

In particular, each ξ_j is in $k(\eta)$. This means that there are polynomials u_j, v_j , with coefficients in k , and without common factors of positive degree, such that $v_j(\eta) \neq 0$ and

$$\xi_j = u_j(\eta)/v_j(\eta) \quad (1 \leq j \leq n). \quad (1)$$

Thus $\xi_j v_j(F(\xi)) - u_j(F(\xi)) = 0$. Property (*) implies now that

$$z_j v_j(F(z)) = u_j(F(z)) \quad (1 \leq j \leq n, z \in \mathbb{C}^n). \quad (2)$$

Put $\Omega = F(\mathbb{C}^n)$. We saw, as a consequence of Fact 2, that Ω is open. If v_j had a zero in Ω , Fact 1 would imply that there is a point in Ω where $v_j = 0$ but $u_j \neq 0$, contradicting (2).

Hence $v_j \circ F : \mathbb{C}^n \rightarrow \mathbb{C}$ is a polynomial without zeros, hence is constant, hence each v_j is constant. Without loss of generality, $v_j = 1$. Putting

$$G = (u_1, \dots, u_n), \quad (3)$$

(2) becomes

$$G(F(z)) = z \quad \text{for all } z \text{ in } \mathbb{C}^n. \quad (4)$$

Hence $F(G(F(z))) = F(z)$. This says that $F \circ G$ is the identity map on Ω . If two polynomials agree on Ω , they agree on \mathbb{C}^n . Thus

$$F(G(w)) = w \quad \text{for all } w \text{ in } \mathbb{C}^n. \quad (5)$$

The theorem follows from (4) and (5), with $F^{-1} = G$.

Lemma. Suppose that \mathcal{F} is a subfield of \mathbb{C} , ξ_1, \dots, ξ_m are in \mathbb{C} , and $\mathcal{F}_1 = \mathcal{F}(\xi_1, \dots, \xi_m)$. Then either $\mathcal{F}_1 = \mathcal{F}$, or there is an isomorphism φ of \mathcal{F}_1 into \mathbb{C} that fixes every element of \mathcal{F} but moves at least one ξ_i .

Proof: Assume $\mathcal{F}_1 \neq \mathcal{F}$. Then there is a nonempty subset of $\{\xi_1, \dots, \xi_m\}$, say (ξ_1, \dots, ξ_j) (after reordering) that is minimal with respect to the property

$$\mathcal{F}_1 = \mathcal{F}(\xi_1, \dots, \xi_j).$$

Put $\mathcal{F}_2 = \mathcal{F}(\xi_1, \dots, \xi_{j-1})$. (This is \mathcal{F} when $j = 1$.) Then

$$\mathcal{F} \subset \mathcal{F}_2 \subsetneq \mathcal{F}_2(\xi_j) = \mathcal{F}_1.$$

Let φ fix every element of \mathcal{F}_2 and choose $\varphi(\xi_j)$ as follows:

If ξ_j is transcendental over \mathcal{F}_2 , let $\varphi(\xi_j)$ be any complex number $\neq \xi_j$ that is also transcendental over \mathcal{F}_2 (such as $1 + \xi_j$).

If ξ_j is algebraic over \mathcal{F}_2 , with minimal polynomial $p(x)$, let $\varphi(\xi_j)$ be another root of $p(x)$.

To every w in \mathcal{F}_1 corresponds a rational function r , with coefficients in \mathcal{F}_2 , such that $w = r(\xi_j)$. Setting $\varphi(w) = r(\varphi(\xi_j))$ gives the desired isomorphism.

REFERENCES

1. J. Ax, A metamathematical approach to some problems in number theory, *Proc. Symp. Pure Math.* vol. 20, (1969), AMS (1971), pp. 161–190.
2. H. Bass, E. J. Connell, D. Wright, The Jacobian conjecture: Reduction of degree and formal expansion of the inverse, *Bull. AMS* 7 (1982), pp. 287–330.
3. A. Bialynicki-Birula and M. Rosenlicht, Injective morphisms of real algebraic varieties, *Proc. AMS* 13 (1962), pp. 200–203.
4. W. Hurewicz and H. Wallman, *Dimension Theory*, Princeton Univ. Press, 1948.
5. O. H. Keller, Ganze Cremona-Transformationen, *Monats. Math. Physik* 47 (1939), pp. 299–306.
6. R. Narasimhan, *Several Complex Variables*, Univ. of Chicago Press, 1971.
7. D. J. Newman, One-one polynomial maps, *Proc. AMS* 11 (1960), pp. 867–870.
8. S. Pinchuk, A counterexample to the real Jacobian conjecture, Preprint, May 1994.
9. J.-P. Rosay, Injective holomorphic mappings, *Amer. Math. Monthly*, 89 (1982), pp. 587–588.
10. W. Rudin, *Principles of Mathematical Analysis*, 3rd Ed., McGraw-Hill, 1976.
11. W. Rudin, *Function Theory in Polydiscs*, Benjamin, 1969.

*Department of Mathematics
University of Wisconsin-Madison
Madison, WI 53706-1388*

An Elementary Proof of the Simplicity of the Mathieu Groups M_{11} and M_{23}

Robin J. Chapman

In this note I prove the simplicity of the Mathieu groups of prime degree, M_{11} and M_{23} , using no group theory beyond Sylow's theorems and basic facts about permutation groups. The only facts about the groups M_{11} and M_{23} which are needed are their orders, and the fact that they are transitive permutation groups on 11 and 23 letters respectively. Most textbooks dealing with the Mathieu groups prove the simplicity of M_{11} by more complicated arguments. For instance Rotman [1], uses a lemma of Burnside whose proof lies beyond the scope of introductory courses on group theory.

Let p be a prime number, and G be a subgroup of S_p , the symmetric group of degree p . It is easy to see that $p \mid |G|$ if and only if G is transitive, i.e., if $1 \leq j, k \leq p$ then there is $\sigma \in G$ with $\sigma(j) = k$, for the only elements of order p in S_p are the p -cycles. We shall assume that G is transitive, and by replacing G by a conjugate if necessary we may also assume that G has $P = \langle (12 \cdots p) \rangle$ as a Sylow p -subgroup. Let $n = |G|$, m_G be the number of Sylow p -subgroups of G , and r_G be the index $|N_G(P) : P|$, where $N_G(P)$ is the normalizer of P in G . As all Sylow p -subgroups of G are conjugate in G then

$$n = |G| = |P| |N_G(P) : P| |G : N_G(P)| = pr_G m_G.$$

By Sylow's third theorem $m_G \equiv 1 \pmod{p}$. Also $P \leq N_G(P) \leq N_{S_p}(P)$ and $N_{S_p}(P)$ is the group of all affine transformations modulo p , i.e., the set of maps of the form

$$x \mapsto ax + b \pmod{p}$$

where $p \nmid a$. Hence $|N_{S_p}(P)| = p(p-1)$ and so $r_G = |N_G(P) : P|$ is a factor of $p-1$. It follows that r_G is the least positive residue of n/p modulo p . The following lemma forms the basis of our proof of simplicity.

Lemma 1. *Let G be a transitive subgroup of S_p , and suppose $m_G > 1$. Then $r_G > 1$.*

Proof: Suppose $m_G > 1$ and $r_G = 1$. Then G has exactly $m_G(p-1) = n - m_G$ elements of order p . Each of these elements has no fixed points on $\{1, 2, \dots, p\}$. Hence G has at most m_G elements with fixed points. Each stabilizer G_j of $j \in \{1, 2, \dots, p\}$ in G consists of m_G elements having at least one fixed point. It follows that $G_1 = G_2 = \cdots = G_p$, the set of all elements of G with fixed points. This means that G_1 is trivial and so $m_G = 1$ contrary to hypothesis. \square

We can now prove the simplicity of an interesting class of groups.

Theorem 1. *Let G be a transitive subgroup of S_p , and suppose $|G| = pmr$ where $m > 1$, $m \equiv 1 \pmod{p}$, $r < p$ and r is prime. Then G is simple.*

Proof: We must have $r_G = r$ and $m_G = m$. Let H be a non-trivial normal subgroup of G . It is easy to see that the orbits of H on $\{1, 2, \dots, p\}$ are permuted by G . As G is transitive and H is non-trivial all the orbits of H must have the same size $s > 1$, so $s = p$ and H is transitive. It follows that $P' \leq H$ for some Sylow p -subgroup P' of G . By Sylow's second theorem all Sylow p -subgroups of G are conjugate in G , and so H contains all Sylow p -subgroups of G . Hence $m_H = m$ and $|H| = pmt$ where $t|r$. But $t > 1$ by the Lemma and as r is prime, $t = r$, $H = G$ and G is simple. \square

We recall briefly some facts about the Mathieu groups, M_{11} , M_{12} , M_{22} , M_{23} and M_{24} . These were the first sporadic simple groups to be discovered—by Mathieu in 1861 and 1873—and are most easily defined as automorphism groups of certain combinatorial structures known as Steiner systems. For instance M_{11} is the automorphism group of the (unique) Steiner system of type $S(4, 5, 11)$ —this is a collection of 5-element subsets of an 11-element set X with the property that each 4-element subset of X is contained in exactly one of the sets in the system. Similarly M_{23} is the automorphism group of the (unique) Steiner system of type $S(4, 7, 23)$. For more details see chapter nine of Rotman's book [1]. Rotman finds the orders of these groups; in particular $|M_{11}| = 7920 = 2^4 \cdot 3^2 \cdot 5 \cdot 11$ and $|M_{23}| = 10200960 = 2^7 \cdot 3^2 \cdot 5 \cdot 7 \cdot 11 \cdot 23$.

Theorem 2. *The Mathieu groups M_{11} and M_{23} are simple.*

Proof: The group M_{11} is a transitive subgroup of S_{11} of order $n = 7920$. Now $n/p = 720 \equiv 5 \pmod{11}$ so $r_G = 5$ and $m_G = 144 > 1$. By Theorem 1 M_{11} is simple.

Similarly the group M_{23} is a transitive subgroup of S_{23} of order $n = 10200960$. Now $n/p = 443520 \equiv 11 \pmod{23}$ so $r_G = 11$ and $m_G = 40320 > 1$. By Theorem 1 M_{23} is simple. \square

From the simplicity of M_{11} and M_{23} it is easy to deduce the simplicity of M_{12} and M_{24} (see Corollary 9.22 in [1]).

REFERENCE

1. J. Rotman, *An Introduction to the Theory of Groups*, (3rd ed.), Allyn and Bacon, 1984.

*Department of Mathematics
University of Exeter
EX4 4QE
United Kingdom
rjc@maths.exeter.ac.uk*

UNSOLVED PROBLEMS

Edited by: **Richard Guy & Richard Nowakowski**

In this department the MONTHLY presents easily stated unsolved problems dealing with notions ordinarily encountered in undergraduate mathematics. Each problem should be accompanied by relevant references (if any are known to the author) and by a brief description of known partial or related results. Typescripts should be sent to Richard Guy, Department of Mathematics & Statistics, The University of Calgary, Alberta, Canada T2N 1N4.

Wanted: A Bad Matrix

Gary H. Meisters

1. THE PROBLEM WITH BAD MATRICES. For vectors x in \mathbb{C}^n , let $\text{diag}(x)$ denote the diagonal matrix whose diagonal entries are the components of the vector x . A given $n \times n$ complex matrix A serves as the *kernel matrix* for the matrix-valued bilinear function $\mathcal{B}(A)(x, y) := 3[\text{diag}(Ax)][\text{diag}(Ay)]A$ of the two vector variables x, y . Here's the question: Is there an $n \times n$ matrix A satisfying both of the following conditions?

Cond 1. The matrix $\mathcal{B}(A)(x, x)$ is nilpotent for all x in \mathbb{C}^n . (The matrix A is *admissible*.)

Cond 2. There are distinct vectors x and y in \mathbb{C}^n such that $\mathcal{B}(A)(x, y)(x - y) = (x - y)$. (The matrix A is *odd*.)

Call a square matrix A satisfying both of these conditions a *bad* matrix.

2. REMARKS ON THE GOOD, THE BAD, AND THE UGLY

2.1. Ott-Heinrich Keller's (1939) Jacobian Conjecture [1, 4, 8, 9] states that: If $\det[F'(x)] \equiv 1$ for a polynomial mapping F , then F is bijective with polynomial inverse. It suffices [2] to prove injectivity. It even suffices [3] to prove injectivity for the special "cubic-linear" maps $F(A)(x) := x - H(A)(x) = x - [\text{diag}(Ax)]^2 Ax$. If the mapping $x \mapsto F(A)(x) := x - H(A)(x)$ is injective, call matrix A *good*. We proved in [10, §3.3 page 118] that A is *good* if and only if Cond 2 is *false*: I.e., A is *good* if and only if it is *not odd*. Cond 2 is certainly false if $\mathcal{B}(A)(x, y)$ is *nilpotent* for all x, y (in which case we say the matrix A is *beautiful*); for then all eigenvalues of $\mathcal{B}(A)(x, y)$ must be *zero* for all x, y . Cond 1 (*admissibility*) is equivalent to $\det[F'(A)(x)] \equiv 1$, which is necessary for $F(A)$ to be injective. See [10, Lemma

1(c) page 112 and Eq. (2.2) page 110]. Thus every beautiful matrix is good, and every good matrix is admissible. The question (Keller's rephrased): Are there any *admissible* matrices that are *odd* (hence *bad*)?

2.2. The matrix-valued bilinear function $\mathcal{B}(A)(x, y)$ has the following properties:

- (a) $\mathcal{B}(A)(x, y) = \mathcal{B}(A)(y, x)$ for all vectors x, y ;
- (b) $\mathcal{B}(A)(x, y)z = \mathcal{B}(A)(x, z)y$ for all vectors x, y, z ;
- (c) $\mathcal{B}(A)(x, x)$ is the Jacobian matrix $H(A)(x)$ of the cubic-homogeneous mapping $H(A)(x) := [\text{diag}(Ax)]^3 \mathbf{1}$, where $\mathbf{1}$ denotes the column $[1, 1, \dots, 1]^T$.

2.3. Here is a 15×15 matrix

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & 2 & -2 & -2 & -2 & 0 & 0 & 2 & 0 & 0 & 2 & 0 \\ 0 & 0 & 2 & 0 & 2 & 0 & -1 & 0 & 0 & -1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & -2 & 4 & 0 & 0 & 0 & -2 & 2 & 2 & 2 & 0 & 0 & 0 & -2 \\ -2 & 0 & -2 & 4 & 0 & 0 & 0 & -2 & 2 & 2 & 2 & 0 & 0 & 0 & -2 \\ 0 & -2 & -2 & 4 & 0 & 0 & 0 & -2 & 2 & 2 & 2 & 0 & 0 & 0 & -2 \\ -2 & 0 & 2 & 0 & 2 & 0 & -1 & 0 & 0 & -1 & 0 & 1 & 1 & 0 & 0 \\ -2 & 0 & 0 & 4 & 2 & -2 & -2 & -2 & 0 & 0 & 2 & 0 & 0 & 2 & 0 \\ 0 & -2 & 0 & 4 & 2 & -2 & -2 & -2 & 0 & 0 & 2 & 0 & 0 & 2 & 0 \\ -2 & 0 & -2 & 0 & -2 & 0 & 1 & 0 & 0 & 1 & 0 & -1 & -1 & 0 & 0 \\ 0 & -2 & 2 & -4 & 0 & 0 & 0 & 2 & -2 & -2 & -2 & 0 & 0 & 0 & 2 \\ 0 & -2 & 0 & -4 & -2 & 2 & 2 & 2 & 0 & 0 & -2 & 0 & 0 & -2 & 0 \\ -2 & -2 & -2 & 4 & 0 & 0 & 0 & -2 & 2 & 2 & 2 & 0 & 0 & 0 & -2 \\ -2 & -2 & 0 & 4 & 2 & -2 & -2 & -2 & 0 & 0 & 2 & 0 & 0 & 2 & 0 \end{bmatrix}$$

which satisfies Cond 1 but not Cond 2. In fact it is good but *ugly* (not beautiful). It has rank 5, nilpotent index 2, and $\mathcal{B}(A)(x, x)^5 = 0$. This example is a slight modification of the example given for another purpose on page 39 of [4]. It is easy to check that $\mathcal{B}(A)(e_1, e_2)$ is not nilpotent (so A is ugly); and that $F(A)$ is injective (so that A is good) [14]. It is harder to show that $\mathcal{B}(A)(x, x)$ has nilpotence-index 5 for all $x \in \mathbb{C}^{15}$. It suffices to show that the 13×13 lower-right block has nilpotence-index 4. This was checked by computer [12].

2.4. All 2×2 admissible matrices A can be written as dyads

$$A = \begin{bmatrix} a \\ b \end{bmatrix} \begin{bmatrix} -b^3 & a^3 \end{bmatrix} = \begin{bmatrix} -ab^3 & a^4 \\ -b^4 & ba^3 \end{bmatrix},$$

for some complex numbers a and b . Furthermore, every such A is beautiful!

2.5. The *cubic-similarity* equivalence relation $A \stackrel{\text{cs}}{\sim} D$: Call matrices A and D *cubic-similar* if there is an invertible matrix P such that $[\text{diag}(APu)]^3 \mathbf{1} = P[\text{diag}(Du)]^3 \mathbf{1} \forall u \in \mathbb{C}^n$. All 2×2 admissible matrices are cubic-similar to the *one* representative $J(1.2) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$. All 3×3 admissible matrices are cubic-similar to one of the *two* representatives

$$J(1.2) = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \text{or} \quad J(2.3) = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}.$$

2.6. We gave six cubic-similarity representatives for 4×4 admissible matrices in [11].

$$\begin{aligned}
 J(1.2) &= \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} & J(2.3) &= \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \\
 N(2.3) &= \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \\
 J(2.2) &= \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} & J(3.4) &= \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \\
 N(3.4) &= \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}
 \end{aligned}$$

They are admissible and mutually inequivalent with respect to cubic-similarity. Integers in their names denote *rank* of A and *nilpotence index* of $\mathcal{B}(A)(x, x)$; both are cubic-similarity invariants [11, 13]. While these and the other representatives shown so far are nilpotent, not every admissible matrix A need itself be nilpotent (see the dyads in 2.4); and not every nilpotent matrix is admissible. However, it follows from a result of Drużkowski [5] that every cubic-similarity equivalence class *contains* a nilpotent matrix. But the ugly 15-dimensional example in Remark 2.3 shows that not every admissible matrix A is cubic-similar to a triangular matrix T , because the matrix $\mathcal{B}(T)(x, y)$ determined by a triangular matrix T is itself triangular, hence nilpotent for all x and y (so both T and A would be beautiful). Admissibility, goodness, and beauty are all cubic-similarity invariants [13]. A computer-check by Engelbert Hubbers [7] verified that all 4×4 admissible matrices are cubic-similar to one of the above *six* representatives.

2.7. Here are *four equivalent formulations* of the cubic-similarity equivalence relation $A \sim^{\text{cs}} D$: There is an invertible matrix P such that for all vectors u, v in \mathbb{C}^n

- (a) $[\text{diag}(APu)]^3 \mathbf{1} = P[\text{diag}(Du)]^3 \mathbf{1}$
- (b) $[\text{diag}(APu)]^2 AP = P[\text{diag}(Du)]^2 D$
- (c) $P^{-1}[H(A)(Pu)]P = H(D)(u)$
- (d) $P^{-1}[\mathcal{B}(A)(Pu, Pv)]P = \mathcal{B}(D)(u, v)$

Note that $[\text{diag}(Ax)]^3 \mathbf{1} = [\text{diag}(Ax)]^2 Ax$. Differentiation of (a) gets (b); and multiplication of (b) on the right by the vector u retrieves (a).

2.8. For the matrix A of 2.3 the characteristic polynomial of $\mathcal{B}(A)(x, y)$ is (from [16])

$$\det[tI - \mathcal{B}(A)(x, y)] = t^{15} + 576(x_1 y_2 - x_2 y_1)^2 t^{13}.$$

It is easily seen from this and the Cayley-Hamilton Theorem that $\mathcal{B}(A)(x, x)$ is nilpotent for all vectors x , and that $\mathcal{B}(A)(x, y)$ is not nilpotent for some distinct x and y . That is, the matrix A of Remark 2.3 is admissible but ugly. This raises more

questions: Which coefficients of the characteristic polynomial of $\mathcal{B}(A)(x, y)$ can be different from zero for admissible matrices A ? (We know, for example, that if A is admissible, then $\det A$ and $\det \mathcal{B}(A)(x, y)$ must both be zero.) What other examples of *admissible-but-ugly* matrices can be found in dimensions $n \geq 5$? What is the smallest dimension containing an ugly matrix?

2.9. Finally, the existence of a real bad matrix A would provide a counterexample to the 1960 *Markus-Yamabe Conjecture* on *global asymptotic stability* described in [9, 10, 15]. Around 1960, Lawrence Markus and Hidehiko Yamabe conjectured that every rest point x_0 of a nonlinear, class \mathcal{C}^1 , n -dimensional system of differential equations $dx/dt = V(x)$ is *globally asymptotically stable* if all the eigenvalues of the Jacobian matrix $V'(x)$ have strictly negative real parts at every point x in \mathbf{R}^n . If there is an $n \times n$ real bad matrix A , then $H(A)(x)^n \equiv 0$ (A is admissible) and the mapping $F(A)(x) \equiv x - H(A)(x)$ is not injective (A is not real good). Thus $F(A)(x_1) = F(A)(x_2)$ for two *distinct* points x_1 and x_2 in \mathbf{R}^n , so that both x_1 and x_2 are rest points of the system

$$\frac{dx}{dt} = V(x) \equiv -x + H(A)(x) + x_1 - H(A)(x_1)$$

because both $V(x_1) = 0$ and $V(x_2) = 0$. Now two distinct rest points cannot *both* be globally asymptotically stable. (A rest point is globally asymptotically stable only if *all* solutions tend to it as “time” t tends to infinity.) However, $H(A)(x)^n \equiv 0$ implies that the Jacobian matrix $V'(x) = -I + H(A)'(x)$ has -1 for all its eigenvalues, so we are in violation of the Markus-Yamabe Conjecture.

3. SUMMARY. We have introduced above the following four classes of complex square matrices A :

1. A is *beautiful* means $\mathcal{B}(A)(x, y)$ is nilpotent for all x, y .
2. A is *not odd* means there are no distinct vectors x, y satisfying

$$\mathcal{B}(A)(x, y)(x - y) = (x - y).$$

3. A is *good* means the map $x \mapsto F(A)(x)$ is injective.
4. A is *admissible* means $\mathcal{B}(A)(x, x)$ is nilpotent for all x .

We know that $(1) \Rightarrow (2) \Leftrightarrow (3) \Rightarrow (4)$. It is also known that (2) does *not* imply (1) . The open question is this: Does (4) imply (2) ? This is Keller’s Question rephrased.

$$\{\text{beautiful}\} \subset \{\text{not odd}\} = \{\text{good}\} \subseteq \{\text{admissible}\}.$$

$$\{\text{bad}\} = \{\text{odd}\} \cap \{\text{admissible}\}.$$

Is $\{\text{bad}\}$ the empty set?

4. ANOTHER QUESTION. Is there a *beautiful* matrix A that is *not cubic-similar to a triangular* (not CST)? It follows from 2.4–2.6 that all 2×2 , 3×3 , and 4×4 admissible matrices are CST. But the 15×15 example in 2.3 is a good matrix that is ugly (hence not CST).

See the references for further details and many related questions. A screenplay written by Luciano Vincenzoni and Sergio Leone inspired our terminology. The author is grateful to the MONTHLY editors and referees for good suggestions that improved the exposition; and to Michael Neubauer who helped him improve it even further.

REFERENCES

1. Hyman Bass, Edwin H. Connell, and David Wright. The Jacobian Conjecture: Reduction of Degree and Formal Expansion of the Inverse. *Bull. A. M. S.* 7(2) (1982), 287–330.
2. Andrzej Białynicki-Birula and Maxwell Rosenlicht. Injective morphisms of real algebraic varieties. *Proc. A. M. S.* 13 (1962), 200–203.
3. Ludwik M. Drużkowski. An Effective Approach to Keller's Jacobian Conjecture. *Math. Ann.* **264** (1983), 303–313.
4. Ludwik M. Drużkowski. The Jacobian Conjecture. IMPAN Preprint 492 (1991), Math. Inst., Jagiellonian University, ul. Reymonta 4, PL-30-059, Kraków, Poland.
5. Ludwik M. Drużkowski. The Jacobian Conjecture in case of rank or corank less than three. *Journal of Pure and Applied Algebra* **85** (1993), 233–244.
6. Arno van den Essen and G. H. Meisters. A Computational Approach to the Jacobian Conjecture. Report 9318, Department of Mathematics, Catholic University, Toernooiveld, 6525 ED Nijmegen, The Netherlands, April 1993.
7. E.-M. G. M. Hubbers, The Jacobian Conjecture: Cubic Homogeneous Maps in Dimension Four. Masters thesis directed by Arno van den Essen at Nijmegen, The Netherlands, February 17, 1994.
8. Ott-Heinrich Keller [22 June 1906–5 December 1990]. Ganze Cremona Transformationen. *Monatshefte für Mathematik und Physik* 47 (1939), 299–306.
Items 6 and 7 in Keller's table on page 301 is the question he raised.
9. G. H. Meisters. Jacobian problems in differential equations and algebraic geometry. *Rocky Mountain J. Math.* 12 (1982), 679–705.
10. G. H. Meisters. Inverting polynomial maps of n -space by solving differential equations. Pages 107–166 in A. M. Fink, R. K. Miller, and W. Kliemann, editors, *Delay and Differential Equations, Proceedings in Honor of George Seifert, Ames, Iowa, October 1991*, World Sci. Pub. Co. Pte. Ltd. Teaneck NJ, 1992. ISBN 981-02-0891-X. (The last sentence on page 151 is false; and lines 6 & 8 page 153.) MR 93g:34072.
11. G. H. Meisters. Power Similarity: Summary of First Results. Conference on Polynomial Automorphisms, held at C.I.R.M. LUMINY, France, October 12–17, 1992.
12. G. H. Meisters. A Good But Ugly Matrix in 15-Dimensions: Its Cubic-Similarity Invariants. A 13.8 MB 5 page *Mathematica* Notebook with 3,117 output pages closed, July 1993, *Mathematica* Version 2.2 for NeXT Computers, Wolfram Research, Inc., 1993. 11 hrs. Appendix: d4pExpanded; a 13 page 5.47MB *Mathematica* Notebook with 2,980 pages closed. 25 hrs. Copies available.
13. G. H. Meisters. Invariants of Cubic-Similarity. In Marco Sabatini, editor, *Recent Results on the Global Asymptotic Stability Jacobian Conjecture*, Dipartimento di Matematica, Università di Trento, I-38050 POVO (TN) ITALY, Sept. 14–17, 1993.
14. G. H. Meisters. Inverting a Cubic-Linear Mapping in 15-Dimensions. A 6 page *Mathematica* Notebook, July 1993, *Mathematica* Version 2.2 for NeXT Computers, Wolfram Research, Inc., 1993. This inversion is also easy by hand. Copies available.
15. G. H. Meisters. The Markus-Yamabe Conjecture Implies the Keller Jacobian Conjecture. In Massimo Furi, editor, *Proceedings on the International Meeting on Ordinary Differential Equations and their Applications* (IMODEA) to celebrate the 70th Birthdays of Roberto Conti and Gaetano Villari, at Firenze, Italy, September 20–24, 1993.
16. G. H. Meisters. The Characteristic Polynomial of $\mathcal{B}(A)(x, y)$ for a Good But Ugly Matrix A in 15-Dimensions. A 9 page 0.163MB *Mathematica* Notebook with 42 pages of intermediate-output closed, January 1994, *Mathematica* Version 2.2 for NeXT Computers, Wolfram Research, Inc., 1993. Computation of the coefficients of the characteristic polynomial of $\mathcal{B}(A)(x, y)$ shows that all but those of t^{15} and t^{13} are zero. Total computing time was approximately 15 hours 10 minutes 50 seconds on a NeXT cube with a 68040 Processor, 64MB RAM, and Version 3.2 NeXTstep System Software. Copies available.
17. G. H. Meisters and Czesław Olech. Power-Exact, Nilpotent, Homogeneous Matrices. *Linear and Multilinear Algebra* 35(3–4) (1993), 225–236.

Department of Mathematics and Statistics
University of Nebraska-Lincoln
Lincoln, NE 68588-0323
meisters@unlinfo.unl.edu

PROBLEMS AND SOLUTIONS

Edited by:
Richard T. Bumby, Fred Kochman and Douglas B. West

Proposed problems should be sent to the MONTHLY PROBLEMS address given on the inside front cover. Please include solutions and relevant references. Three copies of all items needed to evaluate the problem should be sent.

Solutions of published problems should arrive at the MONTHLY PROBLEMS address given on the inside front cover before November 30, 1995. If possible, solutions should be typed with double spacing. Two copies suffice. Several solutions may be mailed together, but they should be on separate sheets of paper. The problem number and the solver's name and mailing address should appear on each solution. A mailing label should be included if an acknowledgment is desired.

The published solution is likely to be based on a solution that is complete and correct. Additional information, such as references to other appearances of the problem or its solution, is also welcome.

An asterisk () after the number of a problem, or part of a problem, indicates that no solution is currently available.*

PROBLEMS

10459. *Proposed by David Beckwith, Sag Harbor, NY.*

A game is played with n disks ($n \geq 3$), each having a black face and a red face. Initially, the n disks are arranged in a circle showing a random pattern of black and red faces. A move consists of taking away a black disk (i.e., one with its black face exposed) and inverting its neighbors (if any). The resulting gap is not closed up, so the remaining disks do not acquire new neighbors. The goal is to remove all the disks. For which initial patterns is this possible?

10460. *Proposed by Torleiv Kløve, University of Bergen, Bergen, Norway.*

Let $(b_1, b_2, \dots, b_{2n})$ be a permutation of $(1, 2, \dots, 2n)$ such that

$(|b_2 - b_1|, |b_3 - b_2|, \dots, |b_{2n} - b_{2n-1}|)$ is a permutation of $(1, 2, \dots, 2n - 1)$.

Show that

$\{b_2, b_4, \dots, b_{2n}\} = \{1, 2, \dots, n\}$ if and only if $b_1 = b_{2n} + n$.

10461. *Proposed by Stephen J. Hershkorn, Rutgers University, New Brunswick, NJ.*

A random variable X is said to be *symmetric* about a real number c if $X - c$ and $c - X$ have the same distribution. Show that X is symmetric about c if and only if

$$\int_c^\infty \Pr(|X - u| \leq a) du = a$$

for all $a > 0$.

10462. *Proposed by Igor Rivin, Melbourne University, Parkville, Victoria, Australia.*

Let Δ and Δ' be nondegenerate simplices in E^n , with $(n - 1)$ -dimensional faces F_i and F'_i respectively ($i = 0, \dots, n$). Let α_{ij} be the dihedral angle between F_i and F_j , and let α'_{ij} the dihedral angle between F'_i and F'_j ($i \neq j$). Prove that if $\alpha_{ij} \geq \alpha'_{ij}$ for all i and j with $0 \leq i < j \leq n$, then Δ and Δ' are similar.

10463. *Proposed by F. S. Cater, Portland State University, Portland, OR.*

Let F_0 be a subfield of the field F_1 . Let X be a (possibly infinite) set and let S be a subset of the vector space F_0^X of all functions from X to F_0 . If S is linearly dependent over F_0 , then S is also linearly dependent over F_1 as a subset of F_1^X . Prove or disprove the converse statement: if S is linearly dependent over F_1 , then S must be linearly dependent over F_0 .

10464. *Proposed by Hillel Gauchman, Eastern Illinois University, Charleston, IL, and Lee A. Rubel, University of Illinois, Urbana, IL.*

Let $z = e^{xy}$, and let \mathbf{W}_n be the n by n matrix whose (i, j) entry (for $0 \leq i, j < n$) is

$$\frac{\partial^{i+j} z}{\partial x^i \partial y^j}.$$

Evaluate $\det \mathbf{W}_n$.

10465. *Proposed by Paul K. Stockmeyer, College of William and Mary, Williamsburg, VA.*

As the Minister of Finance of a newly independent country, it is your job to design a new currency: a sequence $d_1 < d_2 < d_3 < \dots$ of positive integers, with $d_1 = 1$ to be the denominations of various coins and bills. Although you are authorized to create an infinite number of denominations, the legislature has passed some laws restricting your choices.

Rule1: *There must be a bound b on the number of items needed for any payment.*

Rule 2: *the “denomination density”, $\lim_{k \rightarrow \infty} k/d_k$ must be zero.*

Rule 3: *repeatedly choosing the largest denomination less than or equal to the amount remaining to be paid (the greedy algorithm) always leads to the use of the minimal number of items to pay any amount.*

Can you design a currency meeting these rules?

SOLUTIONS

Early Returns in a Tied Election

10248 [1992, 781]. *Proposed by Michael B. Handelsman, Erasmus Hall High School, Brooklyn, NY.*

Candidates Smith and Jones are the only two contestants in an election that will be deadlocked when all the votes are counted—each will receive $2n$ of the $4n$ votes cast. The

ballot count is carried out with successive random selections from a single container. After exactly $2n$ votes are tallied, Smith has S votes and Jones has J votes. What is the expected value of $|S - J|$?

Solution I by Víctor Hernández and Ricardo Vélez, Universidad Nacional de Educación a Distancia, Madrid, Spain The answer is $\binom{2n}{n}^2 \binom{4n}{2n}^{-1}$. More generally, suppose that each candidate will receive N of $2N$ votes cast and that when exactly k votes are tallied, Smith has S_k votes and Jones has J_k votes. Let $X_k = |S_k - J_k|$. When $X_k = j$, the leader has $(k + j)/2$ votes, and the trailer has $(k - j)/2$ votes. Among the remaining votes, there are $(2N - k + j)/2$ for the trailer and $(2N - k - j)/2$ for the leader.

This allows us to compute conditional expectations. Given $X_{k-1} = j$, the value of X_k must be $j + 1$ or $j - 1$. If $j > 0$, then

$$\begin{aligned}\mathbf{E}(X_k | X_{k-1} = j) &= (j + 1) \frac{N - (k - 1 + j)/2}{2N - k + 1} + (j - 1) \frac{N - (k - 1 - j)/2}{2N - k + 1} \\ &= j \frac{2N - k}{2N - k + 1},\end{aligned}$$

while $\mathbf{E}(X_k | X_{k-1} = 0) = 1$. We can also compute conditional expectation in the other direction. Since the sequence of votes is random, $\Pr(X_k = j) = \Pr(X_{2N-k} = j)$, and $\Pr(X_{k-1} = i | X_k = j) = \Pr(X_{2N-k+1} = i | X_{2N-k} = j)$. Hence

$$\mathbf{E}(X_{k-1} | X_k = j) = \mathbf{E}(X_{2N-k+1} | X_{2N-k} = j) = j \frac{k-1}{k}.$$

Now let $p_k = \Pr(X_k = 0)$, so $p_k = 0$ if k is odd, and $p_k = \binom{N}{k/2}^2 \binom{2N}{k}^{-1}$ if k is even. Conditioning on X_k for the computation of $\mathbf{E}(X_{k-1})$, we have

$$\begin{aligned}\sum_{j=1}^{k-1} j \Pr(X_{k-1} = j) &= \mathbf{E}(X_{k-1}) = p_k + \frac{k-1}{k} \sum_{j=1}^k j \Pr(X_k = j) \\ &= p_k + \frac{k-1}{k} \mathbf{E}(X_k).\end{aligned}$$

Next we use this and conditioning on X_{k-1} to compute

$$\begin{aligned}\mathbf{E}(X_k) &= p_{k-1} + \frac{2N-k}{2N-k+1} \sum_{j=1}^{k-1} j \Pr(X_{k-1} = j) \\ &= p_k + \frac{2N-k}{2N-k+1} [p_k + \frac{k-1}{k} \mathbf{E}(X_k)].\end{aligned}$$

Solving for $\mathbf{E}(X_k)$, we have $\mathbf{E}(X_k) = \frac{k(2N-k+1)}{2N} p_{k-1} + \frac{k(2N-k)}{2N} p_k$. With the formula for p_k , this yields

$$\mathbf{E}(X_k) = \begin{cases} \frac{k(2N-k)}{2N} \binom{N}{k/2}^2 \binom{2N}{k}^{-1} & k \text{ is even} \\ \frac{k(2N-k+1)}{2N} \binom{N}{(k-1)/2}^2 \binom{2N}{k-1}^{-1} & k \text{ is odd.} \end{cases}$$

To obtain the result stated at the outset, set $k = N = 2n$.

Solution II by Richard Holzsager, The American University, Washington, DC. Start with the easily checked identity

$$(j-i) \binom{N}{i} \binom{N}{j} = N \left[\binom{N-1}{i} \binom{N-1}{j-1} - \binom{N-1}{i-1} \binom{N-1}{j} \right].$$

Then note that

$$\begin{aligned} \mathbf{E}(X_k) &= 2 \sum_{i=0}^{\lfloor k/2 \rfloor} (k-2i) \binom{N}{i} \binom{N}{k-i} / \binom{2N}{k} \\ &= 2 \sum_{i=0}^{\lfloor k/2 \rfloor} N \left[\binom{N-1}{i} \binom{N-1}{k-i-1} - \binom{N-1}{i-1} \binom{N-1}{k-i} \right] / \binom{2N}{k}, \end{aligned}$$

which telescopes to give

$$\mathbf{E}(X_k) = 2N \binom{N-1}{\lfloor k/2 \rfloor} \binom{N-1}{\lfloor k/2 \rfloor - 1} / \binom{2N}{k}.$$

A straightforward manipulation of binomial coefficients shows this formula to be equivalent to the one given in Solution I. Also, when $k = N = 2n$ as in the original statement, Stirling's formula shows that $\mathbf{E}(X_k)$ is asymptotic to $\sqrt{2n/\pi}$.

Editorial comment. None of the other submitted solutions provided as general a result as these, although the proposer did calculate $\mathbf{E}(X_k)$ for k even. However, several other solvers observed the asymptotic result obtained from Stirling's formula. Both R. Daniel Hurwitz and Robert J. Wagner pointed out the similarity of this problem to Problem 436 in *College Math. J.* [1990, 423; 1991, 444] by the same proposer. That problem involved the binomial distribution where this one involved the hypergeometric distribution. Both of the incorrect solutions used the wrong distribution, thereby solving that problem instead of the one posed here.

Solved also by D. M. Bloom, M. Bowron, R. J. Chapman (U. K.), D. A. Darling, Z. Franco, N. N. Gurwell & E. D. Onstott, V. Hernández & R. Vélez (Spain), R. D. Hurwitz, I. Kastanas, P. G. Kirmser, K. McInturff, G. Schillinger, E. Schmeichel, F. Schmidt, G. L. Stanek, M. Vowe (Switzerland), R. J. Wagner, H. Widmer (Switzerland), Anchorage Math Solutions Group, and the proposer. Two incorrect solutions were received.

A Mean Limit

10259 [1992, 873]. *Proposed by Jonathan L. King, University of Florida, Gainesville, FL.*

Let $\langle r_k \rangle$ for $k \in \mathbb{N}$ be defined by $r_0 = 3$ and $r_{k+1} = r_k^2 - 2$. Evaluate

$$\lim_{K \rightarrow \infty} \sqrt[2^K]{\prod_{k=0}^{K-1} r_k}.$$

Solution I by Yan Loi Wong, National University of Singapore, Singapore. The value is $\frac{3+\sqrt{5}}{2}$. Let α be the positive number such that $\cosh \alpha = 3/2$. By using induction and the identity $\cosh 2x = 2 \cosh^2 x - 1$, we have $r_k = 2 \cosh(2^k \alpha)$. From the identity $\sinh 2x = 2 \sinh x \cosh x$, we have

$$\prod_{k=0}^{K-1} r_k = (2 \cosh \alpha)(2 \cosh 2\alpha) \dots (2 \cosh 2^{K-1} \alpha) = \frac{\sinh(2^K \alpha)}{\sinh \alpha}.$$

Hence

$$\lim_{K \rightarrow \infty} \sqrt[2^K]{\prod_{k=0}^{K-1} r_k} = \lim_{K \rightarrow \infty} \left(\frac{\sinh(2^K \alpha)}{\sinh \alpha} \right)^{1/2^K}.$$

By L'Hôpital's Rule, this limit equals $e^\alpha = \frac{3+\sqrt{5}}{2}$.

Solution II by Robin John Chapman, University of Exeter, Exeter, U.K. Letting $\alpha = (3 + \sqrt{5})/2$, we have $r_0 = \alpha + \alpha^{-1}$. The recurrence then yields $r_k = \alpha^{2^k} + \alpha^{-2^k}$, by induction. Since this equals $(\alpha^{2^{k+1}} - \alpha^{-2^{k+1}})/(\alpha^{2^k} - \alpha^{-2^k})$, we have $\prod_{k=0}^{K-1} r_k = (\alpha^{2^K} - \alpha^{-2^K}) / (\alpha - \alpha^{-1})$. The 2^K th root of this is $\alpha^{\left(\frac{1-\alpha^{-2^{K+1}}}{\alpha-\alpha^{-1}}\right)^{1/2^K}}$, which converges to α as $K \rightarrow \infty$.

More generally, if $r_0 \geq 2$ is a real number, then a similar argument proves that the limit is $\left(r_0 + \sqrt{r_0^2 - 4}\right)/2$.

Solution III by Curtis Cooper, Central Missouri State University, Warrensburg, MO. We use the Fibonacci numbers F_n and the Lucas numbers L_n , defined by

$$F_{n+2} = F_{n+1} + F_n \text{ for } n \geq 0; F_0 = 0, F_1 = 1$$

$$L_{n+2} = L_{n+1} + L_n \text{ for } n \geq 0; L_0 = 2, L_1 = 1.$$

With $\alpha = (1 + \sqrt{5})/2$ and $\beta = (1 - \sqrt{5})/2$, we have the classical formulas $F_n = (\alpha^n - \beta^n)/\sqrt{5}$ and $L_n = \alpha^n + \beta^n$, and hence $L_n F_n = F_{2n}$. By induction on k we have $r_k = L_{2^{k+1}}$ and then, by induction on K , $\prod_{k=0}^{K-1} r_k = F_{2^{K+1}}$. Therefore, $\lim_{K \rightarrow \infty} \sqrt[2^K]{\prod_{k=0}^{K-1} r_k} = \alpha^2 = (3 + \sqrt{5})/2$.

Editorial comment. Variations of this problem recently appeared elsewhere. Several solvers mentioned Problem 1393, *Mathematics Magazine* [1993, 127] which in our notation requests a formula for the product $\prod_{k=0}^{K-1} r_k$ for general $r_0 > 2$. Three solutions are presented, but the nice connection to Fibonacci and Lucas numbers in the case $r_0 = 3$ is not evident. Related references include Problem E3036 of this MONTHLY [1987, 789]. Tareq Alnaffouri reported two published solutions to an equivalent problem for $r_0 > 2$ in Problem B-698, *Fibonacci Quarterly* [1992, 369].

The product in Solution III leads to an efficient computation of $F_{2^{K+1}}$, and can be modified to compute other large Fibonacci numbers. See Paul Cull and James L. Holloway, Computing large Fibonacci numbers quickly, *Info. Proc. Letters* 32(1989), 143–149 for more information.

Jonathan Borwein observed that the existence of $\lim(r_n)^{2^{-n}}$ guarantees that the more complicated expression has the same limit. This allows the result to generalize to other recurrences. For example, for $r_0 = A \geq 2$ and $r_{k+1} = r_k^3 - 3r_k$, we obtain

$$\lim_{K \rightarrow \infty} \left(\prod_{k=0}^{K-1} r_k \right)^{1/3^K} = \left(\frac{A + \sqrt{A^2 - 4}}{2} \right)^2.$$

P.-G. Becker and W. Bergweiler, “Transcendancy of local conjugacies in complex dynamics and transcendancy of their values” (submitted) characterizes polynomials $r(x)$ of degree d such that there is an algebraic function $u(z)$, behaving like λz as $z \rightarrow \infty$ with $r(u(z)) = u(z^d)$. This property of $r(x) = x^2 - 2$ with $u(z) = z + z^{-1}$ was used in Solution II.

Solved correctly by 48 readers and the proposer. Three incorrect solutions were received.

An Insufficient Condition for Primality

10268 [1992, 958]. Proposed by Ondrej Šuch (student), Queens University, Kingston, Ontario, Canada.

Define a sequence $\langle a_n \rangle$ for $n \in \mathbb{N}$ by

$$a_0 = 3 \quad a_1 = 0 \quad a_2 = 2$$

$$a_{n+3} = a_{n+1} + a_n \quad (n \in \mathbb{N}).$$

If p is a prime, show that $p|a_p$.

Solution by Anchorage Math Solutions Group, University of Alaska, Anchorage, AK. Since the roots α, β, γ of $r^3 = r + 1$ are distinct, the general solution of the recurrence is $c_1\alpha^n + c_2\beta^n + c_3\gamma^n$. Since

$$\alpha^2 + \beta^2 + \gamma^2 = (\alpha + \beta + \gamma)^2 - 2(\alpha\beta + \beta\gamma + \alpha\gamma) = 0^2 - 2(-1) = 2,$$

we see that $c_1 = c_2 = c_3 = 1$ fits the initial values, and hence $a_n = \alpha^n + \beta^n + \gamma^n$.

Since α, β and γ satisfy $x^3 = x + 1$, we have

$$a_{3n} = (\alpha + 1)^n + (\beta + 1)^n + (\gamma + 1)^n = \sum_{k=0}^n \binom{n}{k} a_k.$$

Similarly, $a_n = \sum_{k=0}^n (-1)^k \binom{n}{k} a_{3k}$. Since $p \mid \binom{p}{k}$ for $1 \leq k \leq p-1$, and $\binom{p}{0} = \binom{p}{p} = 1$, this gives

$$a_p \equiv a_0 - a_{3p} \equiv a_0 - (a_0 + a_p) \equiv -a_p \pmod{p}$$

when p is odd. Since the claim for $p = 2$ is part of the initial conditions, this implies the result.

Editorial comment. The large majority of solvers obtained the general solution and then reduced it to $(\alpha + \beta + \gamma)^n$ modulo p . Some solved the recurrence using generating functions. Explicit formulas for a_n are already known. Allan Pedersen cites “Girard’s formula” from *Encyklopädie der Mathematischen Wissenschaften*, Leipzig, 1989-1904, I B3b, p451, which in this special case gives $a_n = n \sum (i+j-1)/(i!j!)$, taking the sum over $i, j \geq 0$ such that $2i+3j = n$. István Nemes obtained the same formula from “Waring’s formula”, citing Jordan’s *Calculus of Finite Differences* (reprinted by Chelsea, 1965).

Michael W. Vranos noted that the first composite n such that $n|a_n$ is 521^2 ; this was the only example known to the proposer. Using matrix methods, Kurt Foster found that $n|a_n$ also for $n = 821 \cdot 1231 \cdot 6971 = 7045248121$ and $211 \cdot 3571 \cdot 9661 = 7279379941$. For these values of n and positive integer k , $a_{kn} \equiv a_k \pmod{n}$.

H.-J. Seiffert observed that the problem is a special case of a problem in *Fibonacci Quarterly* (31(1993)2, 188) by Paul S. Bruckman; the solution has not yet appeared. L. Van Hamme noted that a theorem in C. Smyth, “A coloring proof of a generalization of Fermat’s little theorem,” this MONTHLY, 93(1986), 469-471, implies the stronger result that $\sum_{d|n} a_d \mu(n/d) \equiv 0 \pmod{n}$, where μ is the Möbius function. Frank Schmidt cites a solution, along with the examples of composite n including those above, by William W. Adams and Daniel Shanks in “Strong primality tests that are not sufficient,” *Mathematics of Computation* 39(1982), 255-300. These authors trace consideration of the problem to R. Perrin in 1899.

Michael Stoll proved more generally that if f is a monic polynomial of degree m with integer coefficients and roots r_1, \dots, r_m , and $a_n = \sum_{j=1}^m r_j^n$, then $p|(a_p - a_1)$ for every prime p . Fermat’s theorem is the case $m = 1$.

Solved by 50 readers and the proposer.

A Characterization of Small Symmetric Groups

10270 [1992, 958]. *Proposed by Marian Deaconescu, University of Timișoara, Timișoara, Romania.*

Prove that a finite group G has the property

$$N_G(H)/C_G(H) \cong \text{Aut}(H)$$

for all subgroups H if and only if G is isomorphic to one of the groups S_n for $n \leq 3$.

Solution I by National Security Agency Problems Group, Fort Meade, MD. Given a finite group X and a prime p , we let $|X|_p$ denote the highest power of p that divides the order of X . Also $\text{Inn}(X)$ denotes the group of inner automorphisms of X and $Z(X)$ denotes the center of X .

Step 1. All Sylow subgroups of G are cyclic of prime order.

Let P be a Sylow subgroup of G , and suppose that $|P|_p = p^e$ for $e \geq 2$. By a theorem of Gaschutz (see Suzuki, *Group Theory*, Theorem 8.14), the index of the group of inner automorphisms of P in the group of all automorphisms of P is divisible by p . By hypothesis, the automorphism group of P is isomorphic to $N_G(P)/C_G(P)$, and so

$$|\text{Aut}(P)|_p = \frac{|N_G(P)|_p}{|C_G(P)|_p} \leq \frac{|N_G(P)|_p}{|Z(P)|} = \frac{|P|_p}{|Z(P)|_p} = |P/Z(P)|_p = |\text{Inn}(P)|_p,$$

contradicting Gaschutz' theorem. Therefore, $e = 1$ as claimed.

Step 2. G does not have any cyclic subgroups of composite order.

Suppose H is a cyclic subgroup of order m . If m is divisible by two odd primes, then $|\text{Aut}(H)| = \phi(m) \equiv 0 \pmod{4}$. By hypothesis, $\text{Aut}(H) \cong N_G(H)/C_G(H)$. Since $|N_G(H)/C_G(H)|$ divides $|G|$, it follows that 4 divides the order of $|G|$, contradicting Step 1. If $m = 2p$ for an odd prime p , then $C_G(H) \supseteq H$ implies that $|N_G(H)/C_G(H)|$ is odd, but $|\text{Aut}(H)| = \phi(m) = p - 1$ is even.

Step 3. $|G| \leq 6$.

If the Sylow subgroups of a finite group G are cyclic, then either G is cyclic or G is metacyclic and is generated by two elements a and b with the defining relations $a^m = 1$, $b^n = 1$, $b^{-1}ab = a^r$, where $mn = |G|$, $\gcd((r-1)n, m) = 1$ and $r^n \equiv 1 \pmod{m}$ (see Hall, *Group Theory*, Theorem 9.4.3). If G is cyclic, then Step 2 implies that G has prime order or is trivial. In either case $\text{Aut}(G) \cong N_G(G)/C_G(G)$ reduces to the trivial group, and so G must be trivial or cyclic of order 2. If G is not cyclic, then m and n must be prime and $|G|$ is a product of two primes. If G is not divisible by any odd prime, then $|G| \leq 2$ by Step 1. If p is the smallest odd prime dividing $|G|$, and P is a p -Sylow subgroup, then $|G|$ is also divisible by $|\text{Aut}(P)| = p - 1$, which forces $p = 3$.

We conclude that $|G| \in \{1, 2, 6\}$. Since G cannot be cyclic of order 6, the remaining possibility for G is S_3 . It is easy to check that S_3 does satisfy the hypothesis, so G must be S_n for some $n \leq 3$, as desired.

Solution II by F. Schmidt, Arlington, VA. As in Step 1 of Solution I, we conclude that $|G|$ is square-free. Also note that the number of automorphisms a cyclic group of order n is the number of positive integers less than n that are relatively prime to n , which is even if $n > 2$.

For any prime p dividing $|G|$, we have a cyclic subgroup H of order p . Since $|\text{Aut}(H)| = p - 1$, we conclude from the hypothesis on G that $p - 1$ divides $|G|$. Now the order of G is confined to the set T of square-free positive integers s such that $p|s$ implies $(p-1)|s$ for each prime p . If $\prod_{i=1}^k p_i \in T$, then also $\prod_{i=1}^{k-1} p_i \in T$, where p_k is the largest prime in the product. Since $2 \cdot 3 \cdot 7 \cdot 43 + 1$ is not prime, we conclude that $T = \{1, 2, 2 \cdot 3, 2 \cdot 3 \cdot 7, 2 \cdot 3 \cdot 7 \cdot 43\}$.

If $|G| \in \{1, 2, 6\}$, then $G \in \{S_1, S_2, S_3\}$, since the cyclic group of order 6 does not satisfy the hypothesis. For the two other possibilities for $|G|$, let p be the largest prime dividing $|G|$, and let P be a p -Sylow subgroup of G . Since the product of the other primes dividing $|G|$ is $p - 1$, $\text{Aut}(P)$ is the cyclic group Z of order $p - 1$. By the hypothesis, we have $\text{Aut}(P) = Z = N_G(P)/C_G(P)$, which implies that $N_G(P) = G$ and $C_G(P) = P$. Hence $G/P = Z$. Therefore G contains an element of order $p - 1$, which generates a cyclic

subgroup H . Now $\text{Aut}(H) = N_G(H)/C_G(H)$ is impossible, since the left side has even order and the right side does not.

Solved also by S. M. Gagola Jr., D. B. Tyler, and the proposer.

Minimal Polynomials for Irrational Numbers

10272 [1992, 958]. *Proposed by J. Marshall Ash and Leonid Krop, DePaul University, Chicago, IL.*

Show that $\sqrt[n]{2} + \sqrt[n]{3}$ is irrational for $n = 2, 3$, and 4 and find the minimal polynomials that these quantities satisfy.

Solution I by National Security Agency Problems Group, Fort Meade, MD. For any integer $n \geq 2$, $\sqrt[n]{2} + \sqrt[n]{3}$ is an algebraic integer. Therefore, if it is also a rational number, it must be a rational integer. However, $\sqrt[n]{2} + \sqrt[n]{3}$ is between 3 and 4 for $n = 2$ and between 2 and 3 for larger n . Hence it cannot be rational.

Now we describe a method for computing the minimal polynomials of the quantities $\sqrt[n]{2} + \sqrt[n]{3}$, for $n \geq 2$. Define sets A_0, \dots, A_{n-1} as follows: A_i consists of the integers $2^j 3^k$ with $j + k \equiv i \pmod{n}$, for $0 \leq j, k \leq n-1$. Denote the elements of A_i by a_{i1}, \dots, a_{in} . For $0 \leq i \leq n-1$, let B_i be the \mathbb{Z} module with basis elements b_{i1}, \dots, b_{in} , where $b_{ij} = \sqrt[n]{a_{ij}}$. Multiplication by $\sqrt[n]{2} + \sqrt[n]{3}$ effects a mapping from B_i to B_{i+1} (subscripts considered modulo n), which can be given by an $n \times n$ matrix C_i . Each matrix C_i contains exactly two nonzero elements in each row and column. The nonzero elements are all one except for a single 2 and a single 3. The product $C = C_{n-1}, \dots, C_0$ is the matrix for mapping B_0 to itself by $(\sqrt[n]{2} + \sqrt[n]{3})^n$. Therefore, the minimal polynomial for $(\sqrt[n]{2} + \sqrt[n]{3})^n$ divides the characteristic polynomial $p(x)$ of the matrix C . Hence the minimal polynomial for $(\sqrt[n]{2} + \sqrt[n]{3})$ divides $p(x^n)$. Using this technique, we find the following minimal polynomials:

$$\begin{aligned} n = 2: & \quad x^4 - 10x^2 + 1 \\ n = 3: & \quad x^9 - 15x^6 - 87x^3 - 125 \\ n = 4: & \quad x^{16} - 20x^{12} - 666x^8 - 3860x^4 + 1 \\ n = 5: & \quad x^{25} - 5x^{20} - 140x^{15} - 460x^{10} + 35x^5 - 1 \end{aligned}$$

For example, when $n = 3$ we have $A_0 = \{1, 12, 18\}$, $A_1 = \{2, 3, 36\}$, $A_2 = \{4, 6, 9\}$, and

$$\begin{array}{cccc} C_0 & C_1 & C_2 & C \\ \begin{pmatrix} 1 & 0 & 3 \\ 1 & 2 & 0 \\ 0 & 1 & 1 \end{pmatrix} & \begin{pmatrix} 1 & 0 & 3 \\ 1 & 1 & 0 \\ 0 & 1 & 2 \end{pmatrix} & \begin{pmatrix} 2 & 0 & 3 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix} & \begin{pmatrix} 5 & 18 & 18 \\ 3 & 5 & 9 \\ 3 & 6 & 5 \end{pmatrix} \end{array}$$

Solution II and generalization by the proposers. We prove that if n_1, a_1 are positive integers such that a_1^{1/n_1} is not an integer, then $\sum_{i=1}^s a_i^{1/n_i}$ is irrational for all choices of positive integers a_2, \dots, a_s and n_2, \dots, n_s . Let $K = \mathbb{Q}[x_1, \dots, x_s, \omega_1, \dots, \omega_s]$ be the finite field extension of the rationals formed by adjoining x_1, \dots, x_s and $\omega_1, \dots, \omega_s$, where x_j is the positive n_j th root of a_j , and $\omega_j = e^{2\pi i/n_j}$. This extension contains all the solutions of the polynomial $\prod_{j=1}^s [(x^{n_j} - a_j)(x^{n_j} - 1)]$ and hence is normal. Thus there exists an automorphism $\sigma: K \rightarrow K$ fixing each rational that also satisfies $\sigma(x_1) \neq x_1$. Applying σ to the equation $x_j^{n_j} = a_j$ shows that $\sigma(x_j)$ is also a solution to the equation $x^{n_j} = a_j$. Hence $\sigma(x_j) = \eta_j x_j$, where $\eta_j = \omega_j^{k_j}$ with $0 \leq k_j \leq n_j - 1$. Furthermore, $k_1 \neq 0$.

Suppose $r = \sum_{j=1}^s x_j$ is rational. Applying σ yields $r = \sum \eta_j x_j$. Moving all terms to the left and taking real parts yields $\sum (1 - \cos(\arg \eta_j)) x_j = 0$. This forces $\arg \eta_j = 0$ for all j , which contradicts $k_1 \neq 0$.

Solved also by S.-J. Bang (Korea), R. J. Chapman (U. K.), T. P. Dence, H. S. Gunaratne (Brunei), Ignotus (Mozambique), D. W. Koster, K.-W. Lau (Hong Kong), O. P. Lossers (The Netherlands), C. Rees, A. Tissier (France), T. Zeanah, and the GCHQ Problem Solving Group (U. K.)

Rencontres and Random Binary Operations

10280 [1993, 76]. *Proposed by Donald E. Knuth, Stanford University, Stanford, CA.*

Define a random binary operation \star on the set $\{1, \dots, n\}$ by choosing every value independently, so that each of the n^{n^2} possible binary operations is equally likely.

(a) Prove that the axiom

$$((x \star x) \star x) \star ((x \star x) \star x) = x$$

holds for $1 \leq x \leq n$ with probability

$$\sum_{k=1}^n \frac{p_{n,k}}{n^{2n-k}}$$

where $p_{n,k}$ is the number of permutations of $\{1, \dots, n\}$ with k fixed elements.

(b) Show that the probability in (a) is asymptotic to $\frac{1}{2}e^{n-1}n!/n^{2n}$ as $n \rightarrow \infty$.

Note: The sum in (a) should start at 0 rather than 1.

Composite solution by Robin J. Chapman, University of Exeter, Exeter, U. K., and Allan Pedersen, Søborg, Denmark.

(a) If \star satisfies the axiom, define $\pi: \{1, \dots, n\} \rightarrow \{1, \dots, n\}$ by $\pi(x) = x \star x$ for x in $\{1, \dots, n\}$. The axiom shows that π is surjective, and thus π is permutation. The axiom may now be written as $\pi(x) \star x = \pi^{-1}(x)$.

Now let π be a permutation of $\{1, \dots, n\}$ with k fixed points. It will suffice to show that the number of binary operations satisfying the axiom, and associated in this way to π , is $n^{n^2-(2n-k)}$. If \star is such a binary operation then the n products $x \star x$ are determined by π , as are the products $\pi(x) \star x$, of which there are $n - k$ not of the form $x \star x$. The other $n^2 - (2n - k)$ products may be chosen arbitrarily, yielding $n^{n^2-(2n-k)}$ possibilities for \star .

(b) It is well known that $p_{n,k} = \binom{n}{k} p_{n-k,0}$ and that $p_{n,0} = n! \sum_{r=0}^n (-1)^r / r!$. Then we have

$$\sum_{k=0}^n p_{n,k} x^k = n! \sum_{r=0}^n \frac{(x-1)^r}{r!},$$

since the coefficient of x^k on the right is

$$n! \sum_{r=0}^n \frac{(-1)^{r-k}}{r!} \binom{r}{k} = \binom{n}{k} (n-k)! \sum_{r=k}^n \frac{(-1)^{r-k}}{(r-k)!} = p_{n,k}.$$

Thus the probability in (a) is

$$\frac{n!}{n^{2n}} \sum_{r=0}^n \frac{(n-1)^r}{r!}.$$

It is known that $\sum_{r=0}^n n^r / r! \sim (1/2)e^n$ as $n \rightarrow \infty$. (See, for example, D. J. Newman, *A Problem Seminar*, Springer (1982), problem 96.) Thus

$$\sum_{r=0}^n \frac{(n-1)^r}{r!} \sim \frac{1}{2}e^{n-1} + \frac{(n-1)^n}{n!} \sim \frac{1}{2}e^{n-1},$$

by Stirling's formula, and the desired asymptotic formula follows.

Editorial comment. Solvers used a variety of methods for asymptotic evaluation of the sum in (b), of which the simplest is the central limit theorem: if X_1, \dots, X_n are independent Poisson random variables with mean 1 and standard deviation 1 then $S_n = X_1 + \dots + X_n$ has a Poisson distribution with mean n and standard deviation \sqrt{n} . Thus

$$e^{-n} \sum_{r=0}^n \frac{n^r}{r!} = P(S_n \leq n) = P((S_n - n)/\sqrt{n} \leq 0).$$

Since the central limit theorem implies that $(S_n - n)/\sqrt{n}$ approaches a standard normal distribution as $n \rightarrow \infty$, this probability approaches $1/2$.

The proposer noted that the probability in part (a) can be expressed in terms of Ramanujan's function $Q(n)$. Using the asymptotics for $Q(n)$ discussed in his book *Fundamental Algorithms*, Addison-Wesley (1968), p. 117, he gave an asymptotic series for this probability:

$$\frac{1}{2en^n} \left(\sqrt{2\pi n} + \frac{10}{3} + \sqrt{\frac{\pi}{72n}} - \frac{53}{135n} + O(n^{-3/2}) \right).$$

Dennis P. Walsh and the proposer showed that in the variation of this problem in which only the $n^{(n^2+n)/2}$ commutative binary operations are considered, the probability that the axiom is satisfied is asymptotic to $(1/2)e^{n-3/2}n!/n^{2n}$.

Solved also by D. Callan, J. A. Grzesik, S. C. Kian (Singapore), O. P. Lossers (The Netherlands), A. D. Melas (Greece), R. Sprugnoli (Italy), D. P. Walsh, A. N. 't Woord (The Netherlands), GCHQ Problem Solving Group (U. K.), Western Maryland College Problems group, and the proposer. Part (a) only solved by D. Beckwith and J. C. Binz (Switzerland).

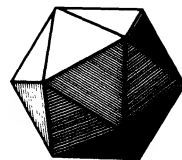
Collaborating editors: David F. Appleyard, Paul T. Bateman, Bruce C. Berndt, Duane M. Broline, Barry W. Brunson, Frank S. Cater, Gulbank D. Chakerian, Underwood Dudley, Gerald A. Edgar, Michael A. Filaseta, Ira M. Gessel, Richard A. Gibbs, Jerrold R. Griggs, Douglas A. Hensley, John R. Isbell, Mourad E. H. Ismail, Murray Klamkin, Daniel J. Kleitman, Frederick W. Luttmann, Frank B. Miles, Richard Pfeifer, Stephen L. Portnoy, J. O. Shallit, John Henry Steelman, Kenneth B. Stolarsky, David E. Tepper, Douglas B. Tyler, Daniel Ullman, and William E. Watkins.

Stirling Numbers

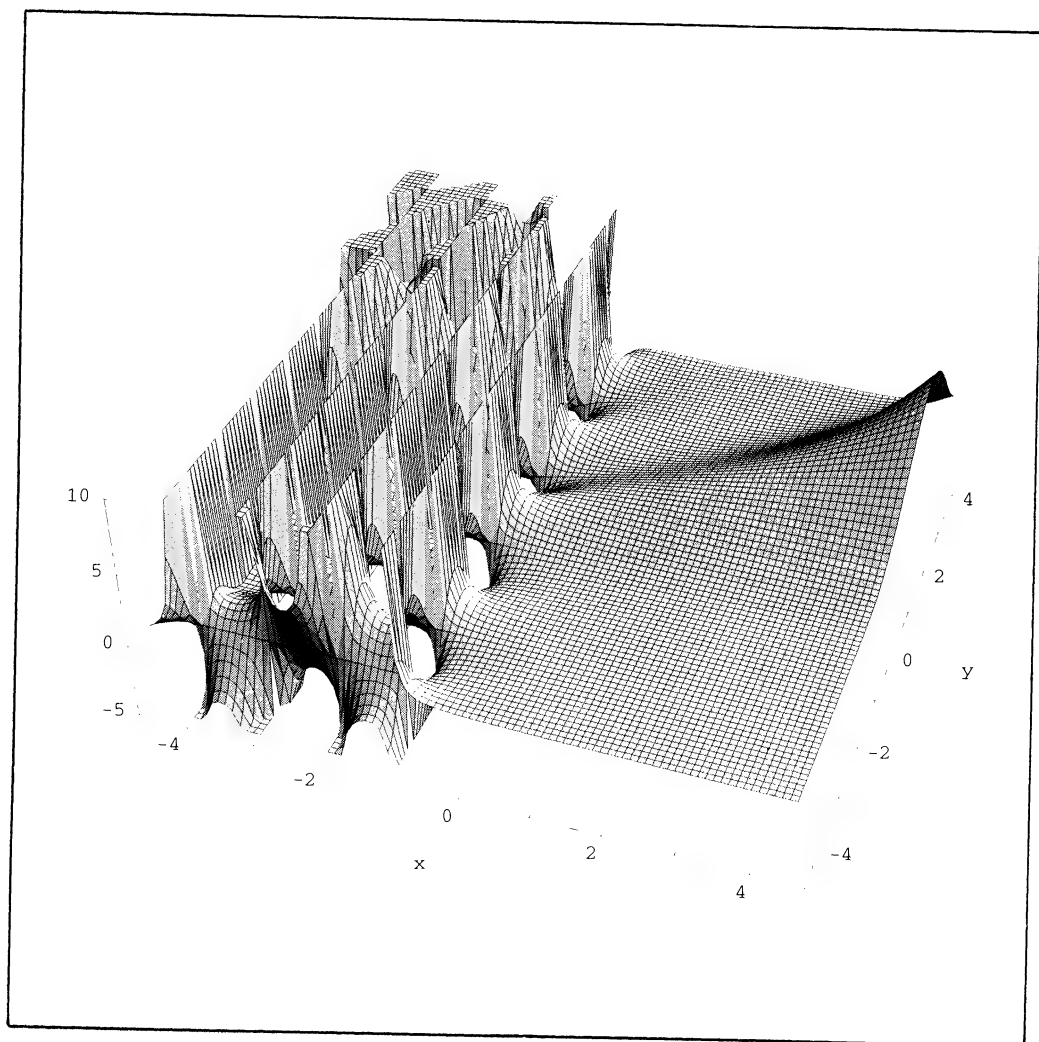
Philippe Flajolet has informed me of the astonishing fact that the notations $\left[\begin{smallmatrix} n \\ k \end{smallmatrix} \right]$ and $\left\{ \begin{smallmatrix} n \\ k \end{smallmatrix} \right\}$ for Stirling numbers recommended in my paper "Two Notes on Notation," *Monthly* **99** (1992), 403–422, were used already by J. Karamata in "Théorèmes sur la sommabilité exponentielle et d'autres sommabilités s'y rattachant," *Mathematica* (Cluj) **9** (1935), 164–178.

Donald E. Knuth
Department of Computer Science
Stanford University
Stanford, CA 94305

The American Mathematical Monthly



Volume 102, Number 7/AUGUST–SEPTEMBER 1995



Have you ever seen this before?
Can you recognize it?
(See page 660)

NOTICE TO AUTHORS

The *Monthly* publishes articles, notes, and other features about mathematics and the profession. The readership of the *Monthly* is intended to include everybody who is mathematically inclined, including of course professional mathematicians and students of mathematics at all collegiate levels. While no single article or feature is likely to appeal to everyone, material should interest and be accessible to a large number of readers. This is the most important criterion for acceptance.

Articles may be expositions of old results or presentations of new ones. They may concern all of mathematics or one small area, a broad development or a single application, historical reminiscences or one important event. While some articles may contain the author's new research, the novelty of material and generality of the results is far less important than the clarity of exposition and general interest. Discussing one illuminating case of a well known result is far better than providing all the details of an obscure but new proposition. Articles in the *Monthly* are supposed to inform and to entertain; they are meant to be read rather than archived.

Notes are short and possibly informal articles. A note may concern a clever new proof of an old theorem, a novel way to present tired material, or a lively discussion of a philosophical (but still mathematical) issue. Also, any topic is suitable, so long as it is related to mathematics. Because a note is short, the first few sentences are the most important part: They should explain the purpose and invite the reader in. Photographs or diagrams often will attract the reader's attention.

All articles and notes should be sent to the editor:

JOHN EWING
Department of Mathematics
Indiana University
Bloomington, IN 47405

Please send 3 copies, typewritten on only one side of the paper. Illustrations should be carefully drawn on separate sheets of paper in black ink; the original should be without lettering and two copies should have appropriate captions and lettering indicated.

Proposed problems or solutions should be sent to:

RICHARD BUMBY,
P.O. Box 10971
New Brunswick, NJ 08906-0971.

Please send 2 copies of all material, typewritten if possible.

Letters to the Editor, both for publication and for private reading, should be sent to the Editor at the address given above. Comments, including criticisms, are welcome, as are all suggestions for making the *Monthly* a lively, entertaining, and informative journal.

EDITOR:

JOHN H. EWING

ASSOCIATE EDITORS:

PETER BORWEIN	FRED KOCHMAN
RICHARD BUMBY	CATHERINE MCGEOCH
DENNIS DETURCK	RICHARD NOWAKOWSKI
UNDERWOOD DUDLEY	ARNOLD OSTEBEE
JOHN DUNCAN	LEE RUBEL
JOAN FERRINI-MUNDY	ABE SHENITZER
JOSEPH GALLIAN	LYNN STEEN
STEVEN GALOVICH	STAN WAGON
RICHARD GUY	DOUGLAS WEST
DARRELL HAILE	HERBERT WILF
PAUL HALMOS	SANDY ZABELL
JOAN HUTCHINSON	PAUL ZORN

EDITORIAL ASSISTANT:

MISTY CUMMINGS

STAFF ARTIST:

MIKE CAGLE

QUOTE MASTER:

MARK WOODARD

Reprint permission:

MARCIA P. SWARD, Executive Director

Advertising Correspondence:

Ms. ELAINE PEDREIRA, Advertising Manager

Subscription correspondence, change of address, and other inquiries:

Membership / Subscriptions Department

All at the address:

The Mathematical Association of America
1529 Eighteenth Street, N.W.
Washington, DC 20036.

Microfilm Editions: University Microfilms International,
Serials coordinator, 300 North Zeeb Road, Ann Arbor, MI 48106.

The AMERICAN MATHEMATICAL MONTHLY (ISSN 0002-9890) is published monthly except bimonthly June-July and August-September by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, DC 20036 and Montpelier, VT. Copyrighted by the Mathematical Association of America (Incorporated), 1995, including rights to this journal issue as a whole and, except where otherwise noted, rights to each individual contribution. General permission is granted to Institutional Members of the MAA for noncommercial reproduction in limited quantities of individual articles (in whole or in part) provided a complete reference is made to the source. Second class postage paid at Washington, DC, and additional mailing offices. **Postmaster:** Send address changes to the American Mathematical Monthly, Membership / Subscription Department, MAA, 1529 Eighteenth Street, N.W., Washington, DC, 20036-1385.

The American Mathematical Monthly

Volume 102, Number 7/AUGUST–SEPTEMBER 1995
(ISSN 0002-9890)



Contents

ARTICLES

Football Pools—A Game for Mathematicians/HEIKKI HÄMÄLÄINEN,
IRO HONKALA, SIMON LITSYN, AND PATRIC ÖSTERGÅRD 579

Fresh Breezes in the Philosophy of Mathematics/REUBEN HERSH 589

Quadratics Representing Primes/NIGEL BOSTON and MARSHALL L.
GREENWOOD 595

How to Write a Proof/LESLIE LAMPORT 600

Searching for Common Generalizations: The Case of Hyperbolic
Functions/KENNETH B. STOLARSKY 609

Transforming n -gons by Folding the Plane/P. SABININ
and M. G. STONE 620

Isometries of the Plane/DAVID A. SINGER 628

FEATURES

COMMENTS 578

NOTES

One More Construction Which Is Impossible/V. A. GEYLER 632

An Inductive Proof of a Mixed Arithmetic-Geometric Mean Inequality
/TAKASHI MATSUDA 634

The Ranks of Tournament Matrices/T. S. MICHAEL 637

THE COMPUTER SCIENCE SAMPLER

On Some Applications of Fibonacci Numbers
/DAVID L. RANUM 640

THE EVOLUTION OF...

Part II. Topology and Abstract Algebra as Two Roads of Mathematical
Comprehension/HERMANN WEYL 646

THE AUTHORS 652

PROBLEMS AND SOLUTIONS 654

REVIEWS

A Radical Approach to Real Analysis by David Bressoud
/SANDY GRABINER 661

TELEGRAPHIC REVIEWS 665

Football Pools—A Game for Mathematicians

Heikki Hämäläinen, Iiro Honkala, Simon Litsyn,
Patric Östergård

1. INTRODUCTION. In a popular game played regularly in many countries one tries to guess the outcomes in a number of competitions or matches. By an outcome we here mean that the host wins, loses or plays a draw, or anyway require that the number of possible outcomes is three (or more generally, some finite number). One can find many small newsletters devoted to this game which contain plenty of different systems for “winning strategies,” that is, sets of guesses which have some nice, relevant properties.

At the same time another group of people—combinatorialists—are busy trying to produce many different types of arrays of numbers with special features. We have the feeling that the two groups are not well acquainted with each other’s work, although recently some mathematical journals have published papers reporting results obtained by the playing community, and we hope that this paper will contribute to increasing the mathematicians’ interest in these problems. The problems are natural and mathematically easy to formulate, but are highly non-trivial and can be attacked using powerful combinatorial machinery.

Some interesting combinatorial objects have been independently discovered by both of these groups. For example, the ternary Golay code was published in the Finnish football pool magazine in 1947 and presented independently in the mathematical literature in 1949 (see [34, Chapter 20] and [16]). Of course, they were also constructed for different purposes: to be used as a particularly nice football pool system by the playing community and as a nice error-correcting code by the mathematical community. Needless to say that the authors were unaware of each other’s discoveries.

In this paper we want to describe several old, well-known problems connected with football pools, as well as some interesting new combinatorial objects arising in this context. We give a rigorous mathematical formulation for each problem and give references to the existing literature and in some cases also tables of the best known numerical results.

To formulate our problem more precisely, assume that we have n matches. In the classical literature one always speaks of football matches, which explains the title of this paper. In each match there are three possible outcomes: 0, 1 and 2. A forecast is a vector (or a word) of length n whose components belong to the set $\mathbf{Z}_3 = \{0, 1, 2\}$, the integers modulo 3. A system (of forecasts) of size M consists of M such vectors. After the matches have been played an entirely correct forecast is said to win the first prize, and more generally, a forecast with $i - 1$ incorrect guesses wins the i th prize. The first prize is usually quite large, and other prizes are smaller and decreasing with i . Furthermore, the size of the i th prize depends

on the total number of forecasts winning the i th prize. Starting from some i , the prizes equal 0 (usually from $i = 5$).

The organizers of this game use a fixed percentage of the stakes to pay the prizes, so without any insight into the teams playing the matches there is really nothing you can do in the long run to get more money than you invested. If you have some expertise, however, you might consider some of the outcomes highly unlikely and wish to exclude them and only concentrate on the remaining possibilities. Then you can use a suitable system and try to win some prize by using as little money as possible. Of course, you always have a chance of a big win as well.

We say that the Hamming distance $d(a, b)$ between two words $a = a_1a_2 \dots a_n$ and $b = b_1b_2 \dots b_n$ is the number of indices i for which $a_i \neq b_i$. The Hamming weight $wt(a)$ is defined to be the number of indices i for which $a_i \neq 0$.

2. PROBLEMS

2.1. The classical football pool problem. The earliest and most natural problem, called *the football pool problem*, is to try to construct a system for m matches which guarantees you at least the second prize. More generally, you wish to construct a system guaranteeing you at least the $(r + 1)$ st prize. Assuming that you already know beforehand (or think that you know) the outcome in some $m - n$ matches reduces the problem to finding a similar system for n matches.

Mathematically, we wish to find the smallest subset S of \mathbf{Z}_3^n such that for every $x \in \mathbf{Z}_3^n$ there exists a word $s \in S$ such that $d(x, s) \leq r$, that is, *the covering radius* of S is at most r . This kind of covering radius problem has been widely studied in information theory [7, 8, 12, 31]. For $r = 0$ the solution is trivial: we simply take $S = \mathbf{Z}_3^n$. For radius $r = 1$ the problem is already open in general.

Example. Consider the case $n = 4$. It can be verified that each of the 81 points of \mathbf{Z}_3^4 is within Hamming distance one from at least one of the nine words

0000, 0112, 0221, 1022, 1101, 1210, 2011, 2120, 2202.

Since each word in \mathbf{Z}_3^n has distance at most 1 to exactly $2n + 1$ words in \mathbf{Z}_3^n we know that when $r = 1$ we need to have at least $3^n/(2n + 1)$ words in our system. This is called *the sphere covering lower bound*:

$$|S| \geq 3^n/(2n + 1).$$

Therefore our system for $n = 4$ is the smallest possible. In general, if n is of the form $(3^h - 1)/2$ then this bound is tight, because of the existence of Hamming codes, cf. [30, p. 36]. The words in the Hamming code are exactly the solutions s of the equation (all the operations are modulo 3)

$$Hs^T = 0,$$

where the columns of H form a maximum set of pairwise linearly independent vectors in \mathbf{Z}_3^h , i.e., they represent distinct points of the projective geometry $PG(h - 1, 3)$.

When $r = 1$ the smallest possible cardinalities of S are also known for the values $n = 1, 2, 3$ and 5, and are 1, 3, 5 and 27, but are unknown for all other values of n .

For the general r , we have the following general *sphere covering bound*:

$$|S| \geq 3^n / \sum_{i=0}^r \binom{n}{i} 2^i.$$

Example. When $1 < r < n$, the sphere covering bound is attained only when $n = 11$ and $r = 2$ [49], [34]. Such a system may be constructed in the following way. Take all the words $(x_1, x_2, \dots, x_{11})$, $x_i \in \mathbf{Z}_3$, $i = 1, \dots, 11$ such that x_1, \dots, x_6 take on all possible combinations of values 0, 1, 2 and the remaining five components are computed from the following system of equations (all the operations modulo 3):

$$\begin{aligned}x_7 &= x_1 + x_3 + 2x_4 + 2x_5 + x_6, \\x_8 &= x_1 + x_2 + x_4 + 2x_5 + 2x_6, \\x_9 &= x_1 + 2x_2 + x_3 + x_5 + 2x_6, \\x_{10} &= x_1 + 2x_2 + 2x_3 + x_4 + x_6, \\x_{11} &= x_1 + x_2 + 2x_3 + 2x_4 + x_5\end{aligned}$$

The constructed set represents the *ternary Golay code*, see [34] and [9] for other definitions and many connections with other areas of mathematics: combinatorics, sphere packings and groups.

In Table 1 the best currently known lower and upper bounds on the smallest possible cardinality $K_3(n, r)$ of a system $S \subseteq \mathbf{Z}_3^n$ with covering radius r are shown.

For constructions of the codes, refinements on the sphere-covering bound for the classical football pool problem, and some earlier results, see, e.g., [1, 2, 3, 4, 11, 13, 16, 23, 24, 25, 26, 27, 31, 36, 37, 39, 43, 50, 51, 52, 54, 56].

TABLE 1. Bounds for $K_3(n, r)$, the minimum cardinality of a set $S \subseteq \mathbf{Z}_3^n$ with covering radius r .

$n \setminus r$	1	2	3
1	1		
2	3	1	
3	5	3	1
4	9	3	3
5	27	8	3
6	63–73	12–17	6
7	150–186	26–34	7–12
8	393–486	52–81	13–27
9	1048–1356	128–219	25–54
10	2818–3645	323–558	57–108
11	7767–9477	729	115–243
12	21395–27702	1919–2187	282–729
13	59049	5062–6561	609–1215

2.2. The binary covering radius problem. As mentioned in the introduction, it is sometimes natural to feel confident that some outcomes do not occur and to exclude them. If we exclude one of the outcomes in each of the n matches we are left with the binary covering radius problem: we wish to find a set $S \subseteq \mathbf{Z}_2^n$ such that every x in \mathbf{Z}_2^n is within Hamming distance r from at least one word $s \in S$. Here \mathbf{Z}_2 is the set of integers modulo 2. In the same way as in the ternary case, we immediately obtain the binary sphere covering bound:

$$|S| \geq 2^n / \sum_{i=0}^r \binom{n}{i}.$$

In fact, if $r = 1$ and n is of the form $n = 2^h - 1$, or $r = 3$ and $n = 23$, or $n = 2r + 1$, or $n = r$, then it can be shown that this bound is tight, and these are the only cases, see [49], [34].

Example. When $n = 4$ and $r = 1$ then each point in \mathbf{Z}_2^4 is within Hamming distance one from at least one of the words
0000, 1000, 0111, 1111.

So, the fact that we decided to exclude one of the possibilities in each match allowed us to decrease the number of words in the system from 9 to 4.

Example [8]. Consider the case $n = 11$ and $r = 1$. The well-known Steiner system $S(4, 5, 11)$ is a collection of 66 5-element subsets called *blocks* of an 11-element set \mathcal{A} and has the property that every 4-element subset of \mathcal{A} is contained in exactly one block. The blocks can be viewed as binary words of length 11, each having exactly five ones, that is, weight five. The 66 blocks and their complements in \mathcal{A} viewed as binary words (of weight 5 and 6) form a system such that every binary word of weight from 4 to 7 is within Hamming distance one from exactly one of these 132 words. What remains is to cover the words of weight from 0 to 2 and their complements in an efficient way. This can be done as follows. Partition the set \mathcal{A} to two parts, one with cardinality five and the other with cardinality six. Take all the 2-elements subsets of these two parts and all the 1-element subsets of the part with five elements. These $10 + 15 + 5 = 30$ subsets and their complements will do. The resulting system contains 192 words and has covering radius one.

In the first column of Table 2 the best currently known lower and upper bounds on the smallest possible cardinality $K(n, 1, 1)$ of a system $S \subseteq \mathbf{Z}_2^n$ with covering radius at most 1 are given.

TABLE 2. Bounds for $K(n, 1, \mu)$, the minimum cardinality of a set $S \subseteq \mathbf{Z}_2^n$ such that every $x \in \mathbf{Z}_2^n$ is within Hamming distance 1 from at least μ elements of S .

n/μ	1	2	3	4
1	1	2		
2	2	3	4	
3	2	4	6	8
4	4	8	11	14
5	7	12	16	22
6	12	19–20	30–32	38–40
7	16	32	48	64
8	32	58–64	90–94	114–125
9	55–62	104–112	154–160	206–220
10	105–120	187–220	289–320	374–416
11	176–192	342–380	512	684–704
12	342–380	631–752	972–1024	1262–1376
13	598–736	1172–1280	1756–1984	2342–2560
14	1171–1408	2186–2560	3356–3776	4370–4992
15	2048	4096	6144	8192
16	4096	7711–8192	11809–12288	15422–16384

For constructions of binary covering codes and methods to derive lower bounds, see [7, 8, 10, 12, 13, 17, 18, 19, 20, 21, 22, 28, 35, 38, 40, 41, 47, 51, 54, 57, 58].

2.3. The mixed case. If we decide to exclude one of the three outcomes only in some, say b , of the n matches, we have the so-called mixed case. The sphere covering lower bound is then

$$|S| \geq 2^b 3^{n-b} \bigg/ \sum_{j=0}^r \sum_{i=0}^j \binom{b}{i} \binom{n-b}{j-i} 2^{j-i}.$$

Example. Consider again the case $n = 4$. Each word in $\mathbf{Z}_3^2\mathbf{Z}_2^2$ is within Hamming distance one from one of the six words (cf. [16, 41]):

0011, 0200, 1000, 1211, 2101, 2110.

So, by excluding one of the three possible outcomes in two of the matches we can decrease the number of words in the system from 9 to 6.

In Table 3 the best currently known upper bounds on the smallest possible cardinality $K_{3,2}(n_1, n_2; 1)$ of a system $S \subseteq \mathbf{Z}_3^{n_1}\mathbf{Z}_2^{n_2}$ with covering radius 1 are given. Lower bounds are not shown. However, the known exact values are marked by a period. For constructions of (these and other) mixed covering codes and lower bounds see [10, 16, 29, 33, 38, 41, 52, 53].

TABLE 3. Bounds for $K_{3,2}(n_1, n_2; 1)$, the minimum cardinality of a set $S \subseteq \mathbf{Z}_3^{n_1}\mathbf{Z}_2^{n_2}$ with covering radius 1.

n_1/n_2	1	2	3	4	5	6	7	8	9	10	11	12
1	2.	3.	6.	8.	16.	24.	48	84	160	284	548	1024
2	4.	6.	12.	20.	36	64	126	234	419	768	1504	
3	9.	16.	24.	48	92	176	320	576	1120	2080		
4	18.	36	72	132	240	432	864	1296	2592			
5	54	96	168	324	639	1206	1944	3888				
6	132	252	468	864	1656	2916	5832					
7	333	648	1296	2304	4374	8640						
8	972	1728	3456	6480	12960							
9	2592	4860	9720	17496								
10	7047	13122	25192									
11	18894	37788										
12	52488											

More generally, we can assume that in each competition there is a different number of possible outcomes. For example, in horse races when trying to pick out the winner in each heat you may have a different number of likely winners that you wish to concentrate on.

2.4. Multiple coverings. Suppose a group of μ players join their efforts and wish to find a system which guarantees them at least μ prizes (one for each!) each of which is at least the $(r + 1)$ st prize. An evident strategy is that each player individually uses a usual football pool system that guarantees at least the $(r + 1)$ st prize. However, this is not always an efficient strategy.

Example. Let $n = 4$ and $r = 1$. As 2-fold and 3-fold coverings of the space \mathbf{Z}_2^4 we could use 2-fold and 3-fold repetitions of the system of four vectors described in the first example of Section 2.2, thus obtaining systems with eight and twelve vectors. Nevertheless, it is possible to do better. The seven words

0001, 0010, 0011, 1100, 1100, 0111, 1011,

and the eleven words

0001, 0010, 0100, 0011, 0101, 1001, 1010, 1100, 0111, 1101, 1110,

provide 2- and 3-fold coverings, respectively. Notice that we have here used the same word more than once, which sometimes does help, cf. [14].

Example. Let $n = 11$ and $r = 1$. We want to find a 3-fold covering of \mathbf{Z}_2^{11} . We construct the system by taking all vectors $(x_1, x_2, \dots, x_{11})$, $x_i \in \mathbf{Z}_2$, $i = 1, \dots, 11$,

such that

$$\begin{aligned}x_{10} &= x_1 + x_2 + x_3 + x_4 + x_5, \\x_{11} &= x_1 + x_2 + x_3 + x_6 + x_7,\end{aligned}$$

(where the summation is taken modulo 2). It is easy to check that the system consists of 2^9 vectors, which is the best possible since it attains the sphere-covering bound:

$$|S| = 2^9 = 3 \times 2^{11}/12.$$

In Table 2 we give the best currently known lower and upper bounds on the smallest cardinality $K(n, 1, \mu)$ of a system $S \subseteq \mathbf{Z}_2^n$ that is a μ -fold covering when $r = 1$. Notice that in Table 2 we have assumed that no word is permitted to appear in the system more than once. For constructions and lower bounds, see [5, 6, 14, 55].

2.5. Multiple coverings of the farthest-off points. Since the first prize is usually much bigger than the second one, and the second one is bigger than the third one, a natural goal is to try to guarantee one big prize or several small prizes.

Example. If we use the system consisting of the six words

$$1111, 1111, 1000, 0100, 0010, 0001,$$

in \mathbf{Z}_2^4 we will always get at least one entirely correct forecast or at least two forecasts with one incorrect entry, as can easily be checked. This means that every word of \mathbf{Z}_2^4 not belonging to the system is covered at least twice. In comparison, we need four words to provide a 1-fold covering and seven words for a 2-fold covering.

For other constructions see [6, 15].

2.6. Weighted coverings with decreasing weights. A general statement of the problem is that the player chooses some decreasing sequence of rational numbers $m = (m_0, m_1, \dots, m_n)$. For every word $x \in \mathbf{Z}_3^{n_1} \mathbf{Z}_2^{n_2}$, $n_1 + n_2 = n$, define the vector $a(x) = (a_0(x), \dots, a_n(x))$, where $a_i(x)$ stands for the number of words in the system being at Hamming distance i from x . A system S is called a weighted m -covering if for every $x \in \mathbf{Z}_3^{n_1} \mathbf{Z}_2^{n_2}$ the inequality

$$\sum_{i=0}^n m_i a_i(x) \geq 1$$

holds.

Notice that the systems in the examples where $n = 4$ are all special cases of this problem for the nonzero weights $m_0 = m_1 = 1, m_0 = m_1 = 1/2, m_0 = m_1 = 1/3$, and $m_0 = 1, m_1 = 1/2$.

Example. Consider \mathbf{Z}_2^4 , and choose $m_0 = m_1 = 1/2, m_2 = 1/4$. The following system of four words

$$0000, 1100, 0011, 1111,$$

will guarantee one first prize and two third prizes, or two second prizes, or four third prizes.

For constructions and bounds, see [6].

2.7. Other related problems. A natural generalization of all the problems mentioned above is to require that only some *part* A of the whole space \mathbf{Z}_3^n is to be taken care of. One obvious choice for A would be that A itself is a Hamming sphere $B_R(x)$ for some integer R and $x \in \mathbf{Z}_3^n$. This sphere consists of the words

within Hamming distance R from the word x . This corresponds to the case where a player is more or less convinced that the outcome x will occur, but accepts the possibility that he may be mistaken in some but not more than R matches. Also, if the player guesses that the number of home wins is at least $n - R$ for some integer R , it is natural to take A as the set of words in which the number of 0's is at least $n - R$, i.e., $B_R(00 \dots 0)$, where $00 \dots 0$ denotes the all-zero word.

Another possible choice for A would be the *complement*

$$\mathbf{Z}_3^n \setminus B_R(x)$$

of a Hamming sphere. This could be useful if a player has seen in a newspaper that the outcome x is the most probable (according to the opinions of experts), and he is personally not at all convinced by their estimate and wishes to cover the area which is as far as possible from the generally accepted guess. In this way he hopes that the number of other people using such forecasts will be small and if he wins a prize he will not have to share it with too many people. There is also another possible motivation for such a choice. Assume that x is the all-two word $22 \dots 2$. Then $\mathbf{Z}_3^n \setminus B_R(x)$ is exactly the set of words in which the number of 2's is smaller than $n - R$. So, if we think that the number of visitor wins is smaller than $n - R$, this seems a reasonable choice. Alternatively, we think that the outcome in each match will be 0 or 1 but we allow for the possibility that we are wrong about some matches.

Example. The best way of covering a Hamming sphere $B_2(000 \dots 0)$ in \mathbf{Z}_2^n , $n \geq 3$, with Hamming spheres of radius 1 uses $n - 1$ spheres: choose the words $000 \dots 0, 111000 \dots 0$ of length n and all the words of weight 1 in \mathbf{Z}_2^n except $1000 \dots 0, 0100 \dots 0, 0010 \dots 0$. Clearly, the $n - 1$ spheres of radius 1 centered at these points cover $B_2(000 \dots 0)$.

To see that this is actually the smallest possible number of such spheres, assume that there are exactly $n - i$ words of weight 1 in our system. Without loss of generality we may assume that these $n - i$ words contain 1's in the last $n - i$ coordinates. Then all the words of weight two that have at least one 1 among the last $n - i$ coordinates are already covered. The number of other words that is required to cover the remaining $\binom{i}{2}$ words of weight 2 is at least $\binom{i}{2}/3$. Hence the total number of words required is at least

$$n - i + \binom{i}{2}/3 > n - 2$$

when $i \leq 2$ or $i \geq 5$. A direct verification shows that the result is correct also when $i = 3$ or $i = 4$.

Example. Consider how $B_3(000 \dots 0) \subseteq \mathbf{Z}_2^n$ can be covered with Hamming spheres of radius 1. Clearly, to cover all the words of weight 3, we must have at least ($n \geq 6$)

$$\binom{n}{3}/(n - 2) = n^2/6 + O(n)$$

spheres of radius 1.

To get an upper bound we use the following construction. We first cover all the words of weight 3 by words of weight 2 by picking to our system all the words of weight 2 in which both the 1's are in even-numbered coordinates or both are in odd-numbered coordinates. Since every word of weight 3 has two 1's in either

odd-numbered or even-numbered coordinates, they are all covered by these words of weight 2. Graph theoretically, we are taking the complement of a triangle-free graph on n vertices (edges correspond to words of weight two). The words of weight 1 are clearly already covered, and to cover all the words of weight 0 and 2, it is sufficient to take the words of weight 1 where the single 1 is in an even-numbered coordinate. All in all, we get the upper bound $n^2/4 + O(n)$.

It is possible to generalize the weighted covering problem even further. Suppose that for each $x \in \mathbf{Z}_3^{n_1} \mathbf{Z}_2^{n_2}$ we have weights $m_i(x)$ depending on x . We now want to find a set $S \subseteq \mathbf{Z}_3^{n_1} \mathbf{Z}_2^{n_2}$ such that

$$\vartheta(x) := \sum_{i=0}^n a_i(x) m_i(x) \geq 1$$

for all $x \in \mathbf{Z}_3^{n_1} \mathbf{Z}_2^{n_2}$. A special case of this is to assume that

$$m_i(x) = m(x) m_i$$

in which case we simply require different densities at different points. From the player's point of view this means that he wishes to take into account some possible outcomes more than others.

REFERENCES

1. A. Blokhuis and C. W. H. Lam, More coverings by rook domains, *J. Combin. Theory Ser. A* 36 (1984), 240–244.
2. W. A. Carnielli, On covering and coloring problems for rook domains, *Discrete Math.* 57 (1985), 9–16.
3. W. A. Carnielli, Hyper-rook domain inequalities, *Stud. Appl. Math.* 82 (1990), 59–69.
4. W. Chen and I. S. Honkala, Lower bounds for q -ary covering codes, *IEEE Trans. Inform. Theory* 36 (1990), 664–671.
5. R. F. Clayton, “Multiple Packings and Coverings in Algebraic Coding Theory,” Ph.D. thesis, Univ. of California Los Angeles, 1987.
6. G. D. Cohen, I. Honkala, S. Litsyn, and H. F. Mattson, Jr., Weighted coverings and packings, submitted for publication.
7. G. D. Cohen, M. G. Karpovsky, H. F. Mattson, Jr., and J. R. Schatz, Covering radius—survey and recent results, *IEEE Trans. Inform. Theory* 31 (1985), 328–343.
8. G. D. Cohen, A. C. Lobstein, and N. J. A. Sloane, Further results on the covering radius of codes, *IEEE Trans. Inform. Theory* 32 (1986), 680–694.
9. J. H. Conway and N. J. A. Sloane, *Sphere-Packings, Lattices and Groups*, Springer-Verlag, New York, 1988.
10. T. Etzion and G. Greenberg, Constructions for perfect mixed codes and other covering codes, *IEEE Trans. Inform. Theory* 39 (1993), 209–214.
11. H. Fernandes and E. Rechtschaffen, The football pool problem for 7 and 8 matches, *J. Combin. Theory Ser. A* 35 (1983), 109–114.
12. R. L. Graham and N. J. A. Sloane, On the covering radius of codes, *IEEE Trans. Inform. Theory* 31 (1985), 385–401.
13. L. Habsieger, Lower bounds for q -ary coverings by spheres of radius one, submitted for publication.
14. H. O. Hämmäläinen, I. S. Honkala, M. K. Kaikkonen, and S. N. Litsyn, Bounds for binary multiple covering codes, *Des. Codes Cryptogr.* 3 (1993), 251–275.
15. H. O. Hämmäläinen, I. S. Honkala, S. N. Litsyn, and P. R. J. Östergård, Bounds for binary codes that are multiple coverings of the farthest-off points, *SIAM J. Discr. Math.*, to appear.
16. H. Hämmäläinen and S. Rankinen, Upper bounds for football pool problems and mixed covering codes, *J. Combin. Theory Ser. A* 56 (1991), 84–95.
17. I. S. Honkala, Modified bounds for covering codes, *IEEE Trans. Inform. Theory* 37 (1991), 351–365.
18. I. S. Honkala and H. O. Hämmäläinen, A new construction for covering codes, *IEEE Trans. Inform. Theory* 34 (1988), 1343–1344.
19. X. Hou, New lower bounds for covering codes, *IEEE Trans. Inform. Theory* 36 (1990), 895–899.

20. X. Hou, An improved sphere covering bound for the codes with $n = 3R + 2$, *IEEE Trans. Inform. Theory* 36 (1990), 1476–1478.
21. J. G. Kalbfleisch and R. G. Stanton, A combinatorial problem in matching, *J. London Math. Soc.* (1) 44 (1969), 60–64; and (2) 1 (1969), 398.
22. J. G. Kalbfleisch and P. H. Weiland, Some new results for the covering problem in W. T. Tutte (ed.), *Recent Progress in Combinatorics*, Academic Press, New York, 1969, pp. 37–45.
23. H. J. L. Kamps and J. H. van Lint, The football pool problem for 5 matches, *J. Combin. Theory* 3 (1967), 315–325.
24. H. J. L. Kamps and J. H. van Lint, A covering problem, in “Colloq. Math. Soc. János Bolyai; Hung. Combin. Theory and Appl.,” Balatonfüred, Hungary, 1969, pp. 679–685.
25. E. Kolev, Codes over GF(3) of length 5, 27 codewords and covering radius 1, submitted for publication.
26. K.-U. Koschnick, A new upper bound for the football pool problem for nine matches, *J. Combin. Theory Ser. A* 62 (1993), 162–167.
27. P. J. M. van Laarhoven, E. H. L. Aarts, J. H. van Lint, and L. T. Wille, New upper bounds for the football pool problem for 6, 7 and 8 matches, *J. Combin. Theory Ser. A* 52 (1989), 304–312.
28. D. Li and W. Chen, New lower bounds for binary covering codes, submitted for publication.
29. B. Lindström, Group partitions and mixed perfect codes, *Canad. Math. Bull.* 18 (1975), 57–60.
30. J. H. van Lint, *Introduction to Coding Theory*, Springer-Verlag, New York, 1982.
31. J. H. van Lint, Recent results on covering problems, in T. Mora (ed.), “Applied Algebra, Algebraic Algorithms and Error-Correcting Codes,” LNCS 357, Springer-Verlag, Berlin, 1989 pp. 7–21.
32. J. H. van Lint, Jr., “Covering Radius Problems,” M.Sc. thesis, Eindhoven University of Technology, The Netherlands, June 1988.
33. J. H. van Lint, Jr. and G. J. M. van Wee, Generalized bounds on binary/ternary mixed packing- and covering codes, *J. Combin. Theory Ser. A* 57 (1991), 130–143.
34. F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*, North-Holland, Amsterdam, 1977.
35. P. R. J. Östergård, A new binary code of length 10 and covering radius 1, *IEEE Trans. Inform. Theory* 37 (1991), 179–180.
36. P. R. J. Östergård, Upper bounds for q -ary covering codes, *IEEE Trans. Inform. Theory* 37 (1991), 660–664; and 37 (1991), 1738.
37. P. R. J. Östergård, Further results on (k, t) -subnormal covering codes, *IEEE Trans. Inform. Theory* 38 (1992), 206–210.
38. P. R. J. Östergård, Construction methods for mixed covering codes, in M. Gyllenberg and L. E. Persson (eds.), *Proceedings of the 21st Nordic Congress of Mathematicians*, Marcel Dekker, New York, 1994.
39. P. R. J. Östergård, New upper bounds for the football pool problem for 11 and 12 matches, *J. Combin. Theory Ser. A*, to appear.
40. P. R. J. Östergård, New constructions for q -ary covering codes, submitted for publication.
41. P. R. J. Östergård and H. O. Härmäläinen, A new table of binary/ternary mixed covering codes, submitted for publication.
42. V. Pless, “Introduction to the Theory of Error-Correcting Codes,” Wiley, New York, 1989.
43. E. R. Rodemich, Coverings by rook domains, *J. Combin. Theory* 9 (1970), 117–128.
44. R. G. Stanton, Covering theorems in groups (or: How to win at football pools), in W. T. Tutte (ed.), *Recent Progress in Combinatorics*, Academic Press, New York, 1969, pp. 21–36.
45. R. G. Stanton, J. D. Horton, and J. G. Kalbfleisch, Covering theorems for vectors with special reference to the case of four and five components, *J. London Math. Soc.* 1 (1969), 493–499.
46. R. G. Stanton and J. G. Kalbfleisch, Covering problems for dichotomized matchings, *Aequationes Math.* 1 (1968), 94–103.
47. R. G. Stanton and J. G. Kalbfleisch, Intersection inequalities for the covering problem, *SIAM J. Appl. Math.* 17 (1969), 1311–1316.
48. O. Taussky and J. Todd, Covering theorems for groups, *Ann. Soc. Polon. Math.* 21 (1948), 303–305.
49. A. Tietäväinen, On the nonexistence of perfect codes over finite fields, *SIAM J. Appl. Math.* 24 (1973), 88–96.
50. E. W. Weber, On the football pool problem for 6 matches: a new upper bound, *J. Combin. Theory Ser. A* 35 (1983), 106–108.
51. G. J. M. van Wee, Improved sphere bounds on the covering radius of codes, *IEEE Trans. Inform. Theory* 34 (1988), 237–245.

52. G. J. M. van Wee, Bounds on packings and coverings by spheres in q -ary and mixed Hamming spaces, *J. Combin. Theory Ser. A* 57 (1991), 117–129.
53. G. J. M. van Wee, On the non-existence of certain perfect mixed codes, *Discrete Math.* 87 (1991), 323–326.
54. G. J. M. van Wee, Some new lower bounds for binary and ternary covering codes, *IEEE Trans. Inform. Theory* 39 (1993), 1422–1424.
55. G. J. M. van Wee, G. D. Cohen, and S. N. Litsyn, A note on perfect multiple coverings of the Hamming space, *IEEE Trans. Inform. Theory* 37 (1991), 678–682.
56. L. T. Wille, The football pool problem for 6 matches: a new upper bound obtained by simulated annealing, *J. Combin. Theory Ser. A* 45 (1987), 171–177.
57. Z. Zhang, Linear inequalities for covering codes: Part I—pair covering inequalities, *IEEE Trans. Inform. Theory* 37 (1991), 573–582.
58. Z. Zhang and C. Lo, Linear inequalities for covering codes: Part II—triple covering inequalities, *IEEE Trans. Inform. Theory* 38 (1992), 1648–1662.

Hämäläinen:
Lehtotie 28
41120 Puuppola
Finland

Honkala:
Department of Mathematics
University of Turku
20500 Turku 50, Finland
honkala@sara.cc.utu.fi

Litsyn:
Department of Electrical Engineering-Systems
Tel-Aviv University
Ramat-Aviv, 69978, Israel
litsyn@eng.tau.ac.il

Östergård:
Department of Computer Science
Helsinki University of Technology
02150 Espoo, Finland
Patric.Ostergard@hut.fi

In mathematics, as in any scientific research, we find two tendencies present. On the one hand, the tendency toward *abstraction* seeks to crystallize the *logical* relations inherent in the maze of materials that is being studied, and to correlate the material in a systematic and orderly manner. On the other hand, the tendency toward *intuitive understanding* fosters a more immediate grasp of the objects one studies, a live *rapport* with them, so to speak, which stresses the concrete meaning of their relations.

—Hilbert

Fresh Breezes in the Philosophy of Mathematics[†]

Reuben Hersh

Since Pythagoras, philosophy of mathematics tried to account for mathematical existence and the nature of mathematical objects.

Numbers, circles, n -dimensional manifolds, all are different from everything else we think about. They're neither physical nor mental. Not mental, because the Pythagorean theorem or any other well-established mathematical fact is independent of what you or I think. Whether we know it and believe it or don't know it and don't believe it, the Pythagorean theorem is still true. Yet it's not physical either! Plato and Aristotle explained that the triangles and circles of the geometer are not physical triangles or circles, but something "ideal."

Spiritual, empirical, psychological, formalist, and logicist explanations have been offered. None give a credible account of what we do when we do mathematics. Presently some authors are constructing a humanist answer.

An Israeli mathematics education researcher, Anna Sfard, recently found an interesting insight. In learning a mathematical concept, children first learn it as algorithm—procedure, or method. Later, the algorithm is transformed into an object. She calls this "reification." It's difficult to achieve, often needing help from teacher. This story is close to theories of the Russian psychologist, Lev Vygotsky.

For example, subtraction is an algorithm. It isn't hard. It reifies into negative numbers—very hard!

Which mathematical entities are frozen algorithms? What's the interaction between doing and being, algorithm and entity? *This is a question in philosophy of mathematics based on mathematical practise, on seeing mathematics as a human activity. It's not a foundationist question.*

FOUNDATIONS LOST. In books on philosophy of mathematics (Korner, or Benacerraf & Putnam) you read of the leading problem, "foundations." How can we establish mathematical knowledge as certain, indubitable, free of any possible doubt? Three historically important solutions to this problem were logicism (Platonism), formalism, intuitionism. All were unsuccessful. For logicism and formalism, no major new idea has come up in over half a century. Intuitionism and its daughter constructivism did strive to carry out the program of Brouwer streamlined by Bishop. But their goal of remaking mathematics constructively is more remote today than 60 or 70 years ago.

The surviving scrap of foundationalism was named "neo-Fregeanism" by Philip Kitcher. This notion still dominates the philosophy of mathematics. It says:

[†]This article originated as an invited talk to the 1993 annual joint meeting of the sections on mathematics and on philosophy of the New York Academy of Science. Thanks to Prof. Bruce Chandler and Prof. Harold Edwards for the invitation to the New York Academy. Double thanks to Prof. Hao Wang of Rockefeller University, whose hospitality in the spring of 1993 was generous and inspiring.

“Philosophical thinking about mathematics need not concern itself with anything but sets, and set theory’s twin sister, logic.” But most researchers, users, teachers, historians of mathematics aren’t primarily interested in sets. So philosophers of mathematics ignore mathematics and mathematicians, and mathematicians find nothing of interest in philosophy of mathematics.

Deplorable! The principal problem in philosophy of mathematics left in paralysis for over half a century! Mathematicians and philosophers of mathematics ignorant of each other’s existence! A Harvard philosopher, Hilary Putnam, published a foundationalist paper titled “Mathematics Without Foundations.” Is philosophy of mathematics pointless and unnecessary? Or is it time for a fresh start?

PHIL / M AND PHIL / SCI. One weird phenomenon of modern philosophy is that philosophy of science and philosophy of mathematics are almost disjoint. Authors in philosophy of science rarely refer to philosophy of mathematics, and vice versa. An author who writes on both subjects, in any one article sticks to one or the other. It’s like baseball and football—play one or the other, but not both at the same time.

I like to compare philosophy of mathematics today to philosophy of science in the 30’s and 40’s. That subject was dominated by logical positivists: Rudolf Carnap and his friends of the “Wiener Kreis” (Vienna Circle). As a result of taking Bertrand Russell and Ludwig Wittgenstein too seriously, they believed they knew the correct methodology for scientific work: (1) state the axioms; (2) give correspondence rules between words and physical observables; (3) derive the theory, as Euclid derived geometry, or Mach derived mechanics.

It was noticed after a while that what logical positivists said had little in common with what scientists did or wanted to do. New ideas in philosophy of science came from Karl Popper, Tom Kuhn, Imre Lakatos, Paul Feyerabend. These subversives disagreed with each other. But they all thought philosophers of science could think about what scientists actually do, not bring presuppositions and instructions for scientists to ignore.

Philosophy of mathematics is overdue for its Popper, Kuhn, Lakatos, and Feyerabend. It’s overdue for analysis of what mathematicians actually do, and the philosophical issues therein.

In fact, this turn is taking place. Wittgenstein and Lakatos helped start it. In recent years Michael Polanyi, George Polya, Alfred Renyi, Leslie White, Ray Wilder, Greg Chaitin, Phil Davis, Paul Ernest, Nick Goodman, Phil Kitcher, Penelope Maddy, Michael Resnik, Gian-Carlo Rota, Brian Rotman, Gabriel Stolzenberg, Robert Thomas, Tom Tymoczko, Jean Paul van Bendegem, and Hao Wang have participated.

Here are ideas some of these people hold.

1) Mathematics is human. It’s part of and fits into human culture. (Not Frege’s abstract, timeless, tenseless, objective reality.)

2) Mathematical knowledge is fallible. Like science, mathematics can advance by making mistakes and then correcting and recorrecting them. (This “fallibilism” is brilliantly argued in Lakatos’ *Proofs and Refutations*.)

3) There are different versions of proof or rigor, depending on time, place, and other things. The use of computers in proofs is a nontraditional version of rigor.

4) Empirical evidence, numerical experimentation, probabilistic proof all help us decide what to believe in mathematics. Aristotelian logic isn’t necessarily always the best way of deciding.

5) Mathematical objects are a special variety of social-cultural-historical object. We can tell mathematics from literature or religion. Nevertheless, mathematical objects are shared ideas, like Moby Dick in literature, or the Immaculate Conception in religion.

How do humanists answer the big question, “What’s the nature of mathematical objects?”

The question seems difficult because of a centuries-old assumption in Western philosophy: “In the world there are two kinds of things. What’s not physical is mental; what’s not mental is physical.” When Frege proved that mathematics is neither physical nor mental, he accounted for it by means of a third kind of entity —“abstract objects”—about which he could say nothing except that they’re neither physical nor mental.

Mental is thought, individual consciousness, subjectivity; wishes, fears, perceptions, hopes, desires, private thoughts.

Matter is what takes up space, has weight, can be studied by scientific instruments. Mountains, bugs, the stars, gamma rays.

Is there anything that’s neither mental nor physical? Yes! Sonatas. Poems. Churches. Religions. Diplomas. Armies. Wars. Universities. Academies of science!

Does the New York Academy of Science exist? Undoubtedly. Is it mental? If the Secretary and the President of the Academy died of amnesia, the life of the Academy would continue. The Academy isn’t just somebody’s thoughts! Even if the building were blown up and the trustees moved the Academy to Yonkers, it would go on. Its physical and mental embodiments are necessary, but they’re not *it*. The Academy isn’t just the minds and bodies of anyone. Neither is it just the stones of its building.

What is it? It’s a social institution. The mental and physical aren’t sufficient to describe the New York Academy of Science. Nor are they sufficient to describe most of the things that most concern us. Marriage and divorce, employment, shopping, prices and salaries, war and peace, professional sports and television shows. All have mental and physical aspects, but they aren’t mental or physical entities. They’re social entities.

There are not two but three basic kinds of things in the world.

Now, what about mathematical objects—let’s just say numbers. If everything’s either mental, physical, or social, then what are numbers? We’ve already seen that numbers aren’t mental or physical. By the law of the excluded middle, they must be social. But let’s not be peremptory. Let’s consider it a hypothesis. Is mathematics social-cultural-historical?

Certainly it’s historical. The history of mathematics is a developed subject. Historians have studied mathematics back to the Babylonians. We don’t know the remote origin of mathematics, or the remote origin of writing, speech, religion, or the family. That origin was part of the self-creation of the human race. Archeology, linguistics, genetics, ethnology tell us a little more. Counting and talking both had their human beginnings.

Mathematics is a social entity. Mathematicians never were isolated hermits. Today they’re in academic, government or industrial jobs, paid directly or indirectly by the government.

Srinivasa Ramanujan, the self-taught Indian mathematical genius, worked hard to be recognized by the English mathematics establishment. Once he was invited, he went to England, at a cost to his family, his religious commitment, and his ability to find daily food he could eat. He did so in order to work with mathematicians who understood what he was doing.

In the 16th and 17th centuries, Fermat, Huygens, Leibnitz were assiduous letter writers, constantly trading ideas with colleagues in other cities and other countries.

Today a new result is certified as part of mathematics after experts read it and pronounce it good. We monitor our product. Acceptance by the profession is essential to be recognized or accepted as a mathematician.

The overall content of mathematics and its direction of movement respond to the pressures of society. The militarization of U.S. mathematics in World War II is an example.

Newton's calculus was a tool in his theory of gravitation. His gravitation theory was a response to the need for better understanding of the motions of planets. The motions of planets were important because England was a maritime nation. Navigational methods better than those of Spain and Portugal had cash value for England.

In saying this, I don't underestimate the insistence of pure mathematicians on autonomy.

TAKING THE TEST. To test a philosophy of mathematics, ask it questions:

- (1) What makes mathematics different?
- (2) What is mathematics about?
- (3) Why does mathematics achieve near-universal consensus?
- (4) How do we acquire knowledge of mathematics, apart from proof?
- (5) Why are mathematical results independent of time, place, race, nationality and gender, in spite of the social nature of mathematics?
- (6) Does the infinite exist? If so, how?
- (7) Why does pure mathematics so often become useful?

The humanist approach gives better answers to questions 1 through 5 than the neo-Fregean, the intuitionist-constructivists, or any other proposed philosophy I know of.

Questions 6 and 7 are harder. I don't say humanism answers these questions. But neither does anybody else.

In conclusion, I want to destroy one of the most popular arrows opponents like to shoot at mathematical humanism.

$2 + 2 = 4$, they say, everywhere and always. In fact, $2 + 2 = 4$ before there were human societies, or even human beings. When 2 brontosauruses went to the water hole and met two other brontosauruses, there were four brontosauruses at the water hole. The truths of mathematics are universal, independent not only of individual consciousness but of social consciousness.

This is Platonism, the view that Wittgenstein attacked so fiercely, and the view, let's face it, that most mathematicians accept.

How can a humanist answer?

First of all, "two" plays two roles. It's an adjective and it's a noun. When you say "two brontosauruses," "two" is an adjective. "Two brontosauruses plus two brontosauruses equals four brontosauruses" is a statement about brontosauruses, not about numbers. Even if you say "Two discrete, reasonably permanent, non-interacting objects collected together with two others of the same ilk makes four such objects," you are talking about properties of discrete, reasonably permanent non-interacting objects. That's a statement in elementary physics.

The noun “two,” on the other hand, as everybody since Pythagoras knows, doesn’t name a physically observable thing. It names some abstract or ideal entity. Plato, Descartes, Frege knew that two is an ideal object. They explained what they meant by an ideal object only in negative terms—not mental, not physical. I’m pointing out that these abstract ideal objects are social concepts.

“But,” says the Platonist, “how can you explain the fact that always and everywhere, regardless of time and place, politics or religion, race or sex, $2 + 2$ always equals 4? The only way to account for it is to say it’s an objective truth, which we all recognize because it’s an objective truth. Otherwise, the universal agreement that $2 + 2 = 4$ would be an inexplicable miracle.”

To this I answer, “It’s bad logic to say something must be true because you can’t think of any other explanation. That’s how philosophers used to prove the existence of a Supreme Creator—they couldn’t conceive any other way for there to be a universe.

“You say that because I haven’t got an explanation that satisfies you about the objectivity of mathematics, therefore I must believe in abstract entities whose relation to the physical world is obscure, which number incredibly remote uncountable infinities, and which are apprehended by our mental or physical faculties in a quite unexplained manner.

“I don’t believe in them. You believe in them only by closing your eyes to their absurdity.”

“I’m aware that some social or intersubjective concepts have the rigidity, the reproducibility, of physical science. The reproducibility of a mathematical calculation is comparable only to the reproducibility of a physical measurement or experiment.”

Somebody might ask, “Why does the physical world have attributes which are so consistent, so reproducible? Why is the gravitational constant the same from one day to the next? Why is the speed of light in vacuum so reliable?”

No physicist or philosopher feels obliged to answer such questions. The possibility of a science of physics is something we accept. We start from there, we don’t try to go back of it. Heidegger asked, “Why is there a universe?” I don’t know what progress he made. Not a promising investigation.

As there’s lawfulness and stability in parts of the physical world, there’s lawfulness and stability in parts of the social-conceptual world. I don’t know why this is so. I’m sure it’s a fruitless question, as fruitless as the same question about the physical world.

Study of the lawful, predictable parts of the physical world has a name. That name is “physics.” Study of the lawful, predictable parts of the social-conceptual world has a name. That name is “mathematics.”

REFERENCES

1. Benacerraf, P. and H. Putnam, Ed. *Philosophy of Mathematics*, second edition, Cambridge University Press, 1985.
2. Davis, P. J. and R. Hersh, 1981 *The Mathematical Experience*, Houghton Mifflin Company, Boston.
3. Ernest, P., Ed. *Mathematics Education and Philosophy: An International Perspective*, The Falmer Press, London, 1994.
4. Ernest, P. *The Philosophy of Mathematics Education*. London, The Falmer Press, 1991.
5. Hersh, R., ed. *New Directions in the Philosophy of Mathematics*, *Synthese*, volume 88 no. 2 August 1991.
6. Kitcher, P. *The Nature of Mathematical Knowledge*, Oxford University Press, 1983.

7. Korner, S. *The Philosophy of Mathematics* London Hutchinson 1980.
8. Sfard, A. 1989 "Translation from Operational to Structural Conception: The notion of function revisited" in Vergnaud et al., 1989. Proceedings of PME 13, Paris, C.N.R.S., University Rene Descartes.
9. Tymoczko, T. *New Directions in the Philosophy of Mathematics*, Birkhauser Boston, 1985.
10. Wang, Hao, *Beyond Analytic Philosophy*. A Bradford Book MIT Press, 1988.

Department of Mathematics
University of New Mexico
Albuquerque, NM 87131
rhersh@math.unm.edu

PICTURE PUZZLE
(from the collection of Paul Halmos)



A double threat family.
(See page 619)

Quadratics Representing Primes

Nigel Boston and Marshall L. Greenwood

1. INTRODUCTION. This paper is the result of a collaboration between two people of different worlds, (Boston) a 32-year-old professional trained at Cambridge and Harvard and (Greenwood) a 78-year-old self-trained amateur working without computer or calculator, using excerpts from [11] D. N. Lehmer's complete list of prime numbers (covering from 2 to 10,006,721) and an IBM Table (covering from 2 to 52,004,201). For small primes ($< 100,000$) he used tables from the Handbook of Math Functions.

The collaboration began in November 1993 when a colleague of the first author presented him with a copy of Issue 30 of the second author's mathematics newsletter [6]. It contained nonstandard notation and language created by the second author. As the first author read it, he realized an interesting problem was being attacked by serious methods. At first he thought that a computer attack would produce better results than many years of working by hand had. It turned out, however, that the second author had found the best example for quadratics with relatively small discriminant (see the remarks after the table in Section 3) and that a better example would require an extensive search.

In the following month or so the second author's health declined, and he agreed to make public his private methods, when urged to do so by Prof. Diamond. Some of his methods were known to the first author, others were clever observations culled from years of work. The first author undertook an extensive computer search, partly using the second author's results and methods and partly going beyond them, and the outcome is presented below.

2. A HISTORY. In 1772 Euler [4] noted that the quadratic $x^2 + x + 41$ represents a prime for $x = 0, 1, \dots, 39$. The reason behind this remarkable property was made clear to number theorists by Rabinovitch [13] at the 1912 ICM, when he showed that for prime $n > 0$, $x^2 + x + n$ represents a prime for $x = 0, 1, \dots, n - 2$ if and only if the field $\mathbb{Q}(\sqrt{1 - 4n})$ has class number one. Since $\mathbb{Q}(\sqrt{-m})$ has class number greater than one if $m > 163$ (as shown by Baker [1], and Stark [16] in 1966/7 and almost shown by a German high school teacher, Heegner, [8] in 1952), Euler's example cannot be extended in this direction. It nevertheless does much better than known linear polynomials in having successive prime values, since the only such known that represents primes for $x = 0, \dots, 21$ is $11410337850553 + 4609098694200x$, discovered recently by Paul Pritchard. See [18] for other good examples.

Euler also considered the quadratic $2x^2 + n$ and showed that for $n = 3, 5, 11, 29$ this represents a prime for $x = 0, \dots, n - 1$. In 1974 Hendy [9] proved that this is so for prime $n > 0$ if and only if $\mathbb{Q}(\sqrt{-2n})$ has class number two. Again Euler's result cannot be extended, since Baker [2] and Stark [17] in 1971 classified all such fields of class number two. There are similar results for polynomials of the form $px^2 + px + n$ (see [9]). For more history on the subject, see [3], vol. I, pp. 420–421. Also, see [12] for partial results on quadratics with positive discriminants.

In another direction, several people have noted that there are, for example, more primes of the form $x^2 - 2$ than of the form $x^2 + 1$ in the long run. In fact there are consistently about 35% more. This was clarified by a conjecture of Hardy and Littlewood [7], who suggested that the number of primes $< N$ represented by $x^2 + c$ (c not minus a square) should be asymptotically of the form $C\sqrt{N}/\log N$, where C is an explicit computable constant. For $x^2 - 2$, $C \cong 1.85$, whereas for $x^2 + 1$, $C \cong 1.37$ [15]. A similar conjecture holds for any polynomial $ax^2 + bx + c$ if we assume that $a > 0$, a, b, c are relatively prime, $a + b, c$ are not both even, and the discriminant $D = b^2 - 4ac$ is not a square. (These conditions remove the most obvious obstructions to the quadratic representing primes for large x . It still has not been shown that every such quadratic represents at least one prime.) The constant C is given explicitly in terms of a, b, c and is largest when D is not a square modulo many small primes. The reason for this will become clear later.

In [5] Fung and Williams found polynomials of the form $x^2 + x + c$ with large C . Their best example was $c = 132874279528931$ for which $C \cong 5.09$. (Compare $C \cong 3.32$ for Euler's quadratic.) The corresponding $D = 1 - 4c = -531497118115723$ is not a square modulo each prime from 3 to 179.

3. THE PROBLEM AND RESULTS. The problem attacked by the second author was to find a quadratic polynomial that represents the most distinct primes for $0 \leq x \leq 99$. If we drop the requirement that the primes be distinct, we can easily get 95 primes by taking Euler's $x^2 - x + 41$ and noting that for $n \leq x \leq n + 99$ for any choice of $n = -65, \dots, -59, -39, \dots, -34$ this quadratic represents 95 primes. Thus, for example, $(x - 34)^2 - (x - 34) + 41 = x^2 - 69x + 1231$ represents 95 primes for $x = 0, \dots, 99$. We will call two quadratics *equivalent* if one can be turned into the other by replacing x by $x + n$ or $n - x$ for some n or if one is minus the other.

We list a representative of each of the 20 equivalence classes we have found that give at least 86 distinct primes for $n \leq x \leq n + 99$ for some n . By replacing x by $x + n$ as in the previous paragraph we then obtain quadratics that represent this many distinct primes for $0 \leq x \leq 99$.

No./100	Quadratic	Discriminant	Intervals $[n, n + 99]$
90/100	$41x^2 + 33x - 43321$	7105733	$n = -57$
88/100	$4x^2 + 2x + 41$	$-652 = -2^2 163$	$n = -40, -39$
88/100	$8x^2 + 2x - 1097$	$35108 = 2^2 8777$	$n = -66, -65, -57, \dots, -53$
88/100	$27x^2 + 3x - 601$	$64917 = 3^2 7213$	$n = -58$
88/100	$9x^2 + 3x - 16229$	$584253 = 3^4 7213$	$n = -53, \dots, -47$
88/100	$37x^2 + 23x - 8863$	1312253	$n = -47, \dots, -42$
88/100	$29x^2 + 9x - 22111$	2564957	$n = -60$
88/100	$67x^2 + 45x - 12569$	3370517	$n = -72, -61$
88/100	$73x^2 + 59x - 18541$	5417453	$n = -60, -55, -54, -53, -39$
88/100	$59x^2 + 3x - 30109$	7105733	$n = -35, \dots, -31$
87/100	$2x^2 - 199$	$1592 = 2^2 398$	$n = 0, 1$
87/100	$8x^2 + 6x - 661$	$21188 = 2^2 5297$	$n = -71$
87/100	$17x^2 + 7x - 20351$	1383917	$n = -57, -56$
87/100	$31x^2 + 21x - 13679$	1696637	$n = -68, -66, \dots, -61$
87/100	$41x^2 + 19x - 29879$	4900517	$n = -40, -39, -36, -35$
87/100	$41x^2 + 39x - 33829$	5549477	$n = -67, -66$
86/100	$x^2 + x + 41$	-163	$n = 0, 1, 2$
86/100	$58x^2 + 42x - 15347$	$3562268 = 2^2 890567$	$n = -53, -52$
86/100	$53x^2 + 35x - 26171$	5549477	$n = -70, -67$
86/100	$82x^2 + 46x - 41647$	$13662332 = 2^2 3415583$	$n = -39, -38$

The squarefree part of the discriminant is also presented, because it is relevant in our discussion later. This question had been considered previously by Karst [10]. His best example was $2x^2 - 199$ above. Apart from this, none of his examples makes the above table. This is probably because all his other quadratics are of the form $ax^2 + bx + c$ with $a = b$, negative discriminant, and no prime factors < 53 . Karst's $2x^2 - 199$ is also discussed in [12].

The two examples above with negative discriminant come out of Euler's work. The example with discriminant 35108 above was discovered by the second author in 1991, presented as $-8x^2 + 530x - 7681$ for $0 \leq x \leq 99$. The example with discriminant 64917 was discovered by the second author in 1992, with a better interval provided by the first author. Our best example translates into $41x^2 - 4641x + 88007$ for $0 \leq x \leq 99$.

4. THE METHODS. The second author began by experimentally gathering "starters" that produce a lot of primes. For instance, he noticed that $8x^2 - 2x - n$ for n a prime that is 7 or $9 \pmod{10}$ and $2 \pmod{3}$ works well in practice. This implies that its discriminant, $4 + 32n$, is $2 \pmod{3}$ and 2 or $3 \pmod{5}$, i.e. not a square modulo these primes.

Each equivalence class of quadratics contains a representative of the form $ax^2 + bx + c$ where $a > 0$ and $a \geq b \geq 0$. We can therefore restrict our attention to quadratics of this form. Let $D = b^2 - 4ac$.

If $ax^2 + bx + c$ is going to be prime for many values of x , then it cannot ever be even, because its parity depends only on whether x is even or odd. If x is even, $ax^2 + bx + c \equiv c \pmod{2}$. If x is odd, then $ax^2 + bx + c \equiv a + b + c \pmod{2}$. This is why, for our purposes, we need only look at quadratics such that c is odd and $a + b$ even.

Suppose that p is an odd prime not dividing a . Consider the equation $ax^2 + bx + c \equiv 0 \pmod{p}$. The quadratic formula gives its solutions in terms of \sqrt{D} . It is then easy to see that $ax^2 + bx + c$ is not divisible by p for any x , precisely when D is a quadratic non-residue (i.e. not a square) modulo p . Since $ax^2 + bx + c \pmod{p}$ depends only on $x \pmod{p}$, we therefore need D to be a quadratic non-residue modulo lots of small primes. The first author conducted a search for D that are quadratic non-residues for at least 10 of the 11 primes between 3 and 37. This search extended from $-200,000$ to $2,000,000$. Positive discriminants were preferred because they tend to produce better quadratics. In addition, some of the best discriminants (e.g. $3^2 7213$) were multiplied by small squares and then tested. He later conducted a larger search (from $-1,000,000$ to $60,000,000$), capturing those D with small $\sum_p [100(1+n)/p]$, where $n = 0$ if $D \equiv 0 \pmod{p}$, $n = 1$ if D is a square mod p , and $n = -1$ otherwise. (This sum estimates how many divisibilities by small primes will occur.)

Once a suitable D was found, he let b run from 0 to 100, found all possible corresponding a and c by looking at divisors of $D - b^2$, and then checked each $ax^2 + bx + c$ so produced to see if it tended to yield lots of primes. If so, it then qualified for finer testing to find intervals of length 100 on which it would produce the most distinct primes.

One problem we ran up against was that of dealing with quadratics of the form $f(x) = ax^2 + bx + c$ with a dividing b (so $b = 0$ or a by the restrictions imposed earlier). In these cases, for each x there is another integer y such that $f(x) = f(y)$ and so to avoid repetition we have to restrict the intervals considered. A way around this for quadratics $ax^2 + ax + c$, noticed by the second author, is to consider instead the related quadratic $4ax^2 + 2ax + c$. Note, for example, how

$x^2 + x + 41$ represents at most 86 primes for 100 consecutive x , but the related $4x^2 + 2x + 41$ does 2 better.

5. THE CHALLENGE. So here's the challenge. Can you beat 90/100? In fact can you make 100/100? According to a famous conjecture in number theory you can. This is Schinzel's Hypothesis (H) [14], which says that if $f_1(x), \dots, f_s(x)$ are irreducible polynomials with integer coefficients, such that no integer $n > 1$ divides $f_1(x), \dots, f_s(x)$ for all integers x , then there should exist infinitely many x such that $f_1(x), \dots, f_s(x)$ are simultaneously prime.

Consider, for instance, $f(x) = x^2 + x + 17959429571$ (gleaned from [5]). Let $s = 100$ and $f_i(x) = f(x + i)$ for $i = 1, \dots, 100$. We show that

(*) no integer $n > 1$ divides $f_1(x) \dots f_{100}(x)$ for all x .

Hypothesis (H) then says that there should be infinitely many x for which 100 consecutive values of $f(x)$ are prime. These primes are distinct ($f(x) = f(y)$ if and only if $y = x$ or $-1 - x$, so the 100 values of x would have to straddle 0 for repetition, which they don't). An equivalent quadratic will then be prime for all $x = 0, \dots, 99$.

To prove (*), suppose a prime n divides $f_1(x) \dots f_{100}(x)$ for all x . For a start, $n > 127$, since the discriminant of f , -71837718283 , is not a square modulo all primes between 3 and 127. For $n = 131$, the roots of $f(x) \bmod n$ are 63 and 67. Choosing $x = 68$ then shows that this n does not satisfy our hypothesis. For each prime $n > 131$, either f has no roots mod n or has roots spaced by > 100 . (Note that for primes > 200 this is automatic.)

In practice, such a sequence of 100 values could only occur for astronomically large x . Yihsiang Liow has searched for sequences of consecutive prime values of f for $x < 100,000,000$. No sequence found so far has more than twenty terms.

It is interesting to note that similar congruence conditions show that in any interval of length 100 at least 6 values of our best quadratic $f(x) = 41x^2 + 33x - 43321$ must be composite. Hypothesis (H) would then say that there exists an interval of length 100 containing 94 distinct prime values of $f(x)$. It is no wonder that Hypothesis (H) is doubted by a number of people.

6. RELATED RESULTS. In conclusion, here are a few interesting related results. The last section discussed whether it would be possible to get 100 distinct consecutive prime values of a quadratic. As a more modest aim, consider the question of beating Euler's example $x^2 + x + 41$ which represents 40 distinct primes for consecutive values of x . In late 1988, Gilbert Fung and Russell Ruby did this, their respective examples being $47x^2 - 1701x + 10181$ and $36x^2 - 810x + 2753$ [12]. These give distinct primes for $x = 0, 1, \dots, 42$ and $x = 0, 1, \dots, 44$ respectively.

Another question attacked by the second author (and unbeaten by the efforts on computer of the first author) is to find quadratics that are even half the time but that represent distinct primes for as many of the 50 remaining x as possible. The examples $-4x^2 + 381x - 8524$ and $-2x^2 + 185x - 3181$ each give 48 distinct primes for $0 \leq x \leq 99$. Can you find an example that gives 49 or even 50 distinct primes?

ACKNOWLEDGMENTS. The first author thanks God for leading him to results. He thanks Harold Diamond, Will Galway, and Yihsiang Liow for helpful discussions regarding this work.

REFERENCES

1. A. Baker, Linear forms in the logarithms of algebraic numbers, *Mathematika* 13 (1966), 204–216.
2. A. Baker, Imaginary quadratic fields with class number two, *Ann. of Math.* 94 (1971), 139–152.
3. L. E. Dickson, *History of the theory of numbers*, Chelsea, New York, 1971.
4. L. Euler, Mém. de Berlin, année 1722, 36, *Comm. Arithm.* 1, 584.
5. G. W. Fung and H. C. Williams, Quadratic polynomials which have a high density of prime values, *Math. Comp.* 55 (1990), 345–353.
6. M. L. Greenwood, *Mathematics Newsletter*, 31 issues. Vol. I, Mathematics by a non-mathematician, 245pp handwritten, San Diego Public Library, Main Library, and White Tower Library, Los Angeles.
7. G. H. Hardy and J. E. Littlewood, Partitio numerorum III: On the expression of a number as a sum of primes, *Acta Math.* 44 (1923), 48.
8. K. Heegner, Diophantische Analysis und Modulfunktionen, *Math. Z.* 56 (1952), 227–253.
9. M. D. Hendy, Prime quadratics associated with complex quadratic fields of class number two, *Proc. Amer. Math. Soc.* 43 (1974), 253–260.
10. E. Karst, New quadratic forms with high density of primes, *Elem. d. Math.* 28 (1973), 116–118.
11. D. N. Lehmer, *List of prime numbers* 1 to 10,006,721, Carnegie Institution of Washington, 1914.
12. R. A. Mollin and H. C. Williams, Class number problems for real quadratic fields, Number theory and cryptography, vol. 154, *LMS Lecture Note Series*, 1990.
13. G. Rabinovitch, Eindeutigkeit der Zerlegung in Primzahl faktoren in quadratischen Zahlkörpern, *Fifth Internat. Congress Math.* (Cambridge), vol. I, 1913, pp. 418–421.
14. A. Schinzel and W. Sierpiński, Sur certaines hypothèses concernant les nombres premiers. Remarque, *Acta Arithm.* 4 (1958), 185–208.
15. D. Shanks, On the conjecture of Hardy and Littlewood concerning the number of primes of the form $n^2 + a$, *Math. Comp.* 14 (1960), 321–332.
16. H. M. Stark, A complete determination of the complex quadratic fields of class-number one, *Michigan Math. J.* 14 (1967), 1–27.
17. H. M. Stark, A transcendence theorem for class number problems, *Ann. of Math.* 94 (1971), 153–173.
18. S. A. Weintraub, Consecutive primes in arithmetic progressions, *J. of Math. Rec.* 25, no. 3 (1993), 169–171.

Department of Mathematics
University of Illinois
Urbana, IL 61801
boston@math.uiuc.edu

3945 Alabama Street
San Diego, CA 92104-2701

For runners, it's often frustrating to hear other athletes dismiss their sport as a torturous yet necessary evil that must be confronted if one is going to "get in shape." (Soccer players, basketball players, et al. are notorious for this sort of unwarranted aggression.) The field of mathematics, to draw what I hope is fresh analogy, is akin to running in this very respect: it is a discipline often written off by engineers, biologists, chemists, and other applied science types as a laborious means to an end. What the cynics fail to appreciate is the degree to which a simple, cogent proof or a graceful, powerful stride are things of beauty independent of their utility with respect to other practices. For those who align themselves with the engineers and the soccer players, please read no further! Mathematical induction (or simply "induction" as it shall henceforth be known) is a technique whose subtlety and force you will fail to appreciate.

Ben Rutter (sophomore at Swarthmore)

(Introductory paragraph in a paper explaining induction written for a discrete math course.)

How to Write a Proof

Leslie Lamport

1. MATHEMATICAL PROOFS. Mathematical notation has improved over the past few centuries. In the seventeenth century, a mathematician might have written

There do not exist four positive integers, the last being greater than two, such that the sum of the first two, each raised to the power of the fourth, (1)
equals the third raised to that same power.

How much easier it is to read the modern version

There do not exist positive integers x , y , z , and n , with $n > 2$, such that (2)
 $x^n + y^n = z^n$.

Yet, the structure of mathematical proofs has not changed in 300 years. The proofs in Newton's *Principia* differ in style from those of a modern textbook only by being written in Latin. Proofs are still written like essays, in a stilted form of ordinary prose.

Formulas written in prose, like (1), are hard to understand and hard to get right. Proofs written in prose are also hard to understand and hard to get right. Anecdotal evidence suggests that as many as a third of all papers published in mathematical journals contain mistakes—not just minor errors, but incorrect theorems and proofs.

Statement (2) is easier to read than statement (1) for two reasons: variables are given names, and formulas are written in a more structured fashion. The benefits of using names is obvious. The benefit of structure is less obvious; we are so used to formulas like $x^n + y^n = z^n$ that we tend to take their structure for granted, and to think they are easy to read just because they are short. Although the brevity of the formula helps, it is primarily its structure that makes it easier to understand than a prose version. The expression

x raised to the power n
plus
 y raised to the power n equals z raised to the power n

is quite long, but it is easy to read because of its structure.

The same principles that make formulas easier to understand can make proofs easier to understand: proof steps should be referred to by name, and the structure of the proof should be manifest.

The proof style I advocate is a refinement of one, called *natural deduction*, that has been used by some logicians for almost a century. Natural deduction has been viewed primarily as a method of writing proofs in a formal logic. What I will describe is a practical method for writing the less formal proofs of ordinary mathematics. It is based on hierarchical structuring—a successful tool for managing complexity.

A method for structuring proofs was presented by Leron [5]. However, his goal was to communicate proofs better, not to make them more rigorous. Despite their hierarchical structuring, the proofs Leron advocated are quite different from the ones presented here. They do not seem to be any better than conventional proofs for avoiding errors.

Avoiding mistakes when manipulating formulas requires careful, detailed calculations. Avoiding mistakes when proving theorems requires careful, detailed proofs. When first shown a detailed, structured proof, most mathematicians react: “I don’t want to read all those details; I want to read only the general outline and perhaps some of the more interesting parts.” My response is that this is precisely why they want to read a hierarchically structured proof. The high-level structure provides the general outline, readers can look at as much or as little of the lower-level detail as they want. However, until one gets used to them, structured proofs do look intimidating.

The ideal tool for reading a structured proof would be a computer-based hypertext system. It would allow the reader to concentrate on a particular level in the structure, suppressing lower-level details. In a printed version, one can ignore lower-level details only by skipping over that part of the text. While this is not ideal, the structure is displayed by the format, making such skipping fairly easy—certainly much easier than in a prose-style proof, where the format provides little clue to the logical structure.

2. AN EXAMPLE. I take as an example the classic proof that $\sqrt{2}$ is irrational. Letting \mathbf{Q} denote the set of rationals, the precise statement of the result to be proved is

Theorem. There does not exist r in \mathbf{Q} such that $r^2 = 2$.

To illustrate hierarchical structure, the proof is carried out to a much lower level of detail than necessary for a typical reader.

2.1. The High-Level Proof. The high-level structure of the proof—what one would see first with a hypertext system—appears in Figure 1. The proof assumes a lemma from which one can deduce that, for any integer n , if 2 divides n^2 then 2 divides n . The set of integers is denoted by \mathbf{Z} .

Theorem. There does not exist r in \mathbf{Q} such that $r^2 = 2$.

PROOF SKETCH: We assume $r^2 = 2$ for $r \in \mathbf{Q}$ and obtain a contradiction. Writing $r = m/n$, where m and n have no common divisors (Step 1), we deduce from $(m/n)^2 = 2$ and the lemma that both m and n must be divisible by 2 (Steps 2 and 3).

ASSUME: 1. $r \in \mathbf{Q}$
2. $r^2 = 2$

PROVE: False

1. Choose m, n in \mathbf{Z} such that
 1. $\gcd(m, n) = 1$
 2. $r = (m/n)$
2. 2 divides m .
3. 2 divides n .
4. Q.E.D.

Figure 1. The highest level of a structured proof of the irrationality of $\sqrt{2}$.

After the statement of the theorem comes a **PROOF SKETCH**, which is an informal explanation of the following proof. The proof sketch serves as a “road map” to the proof, helping the reader understand intuitively why the proof works. This proof is so simple that the proof sketch is almost superfluous—the only information it provides that is not obvious from the high-level proof itself is that the lemma is used to prove Steps 2 and 3.

Next comes the **ASSUME** and **PROVE** clauses. They assert that to prove the theorem, it suffices to assume the two hypotheses $r \in \mathbf{Q}$ and $r^2 = 2$, and to prove *false*.

Finally comes the proof. This is a sequence of statements that ends with “Q.E.D.,” which denotes the assertion to be proved—in this case, *false*. Think of this proof as the left half (the statements) of a high-school geometry style proof, the right half (the reasons) being omitted.¹

2.2. Lower Levels of the Proof. Let us now examine the proof of Step 1, which appears in Figure 2. It is clear enough what must be proved, so no **ASSUME** / **PROVE** is needed. The proof consists of five steps, numbered 1.1 through 1.5. There is also a **LET** statement, which defines the required m and n . (I prefer \triangleq to the more common symbol \equiv for “equals by definition,” since \equiv can also mean logical equivalence.)

1. Choose m, n in \mathbf{Z} such that
 1. $\gcd(m, n) = 1$
 2. $r = (m/n)$
- 1.1 Choose p, q in \mathbf{Z} such that $q \neq 0$ and $r = p/q$.
 LET: $m \triangleq p/\gcd(p, q)$
 $n \triangleq q/\gcd(p, q)$
- 1.2. $m, n \in \mathbf{Z}$
- 1.3. $r = m/n$
- 1.4. $\gcd(m, n) = 1$
- 1.5. Q.E.D.

Figure 2. The proof of Step 1.

Each of these five steps in turn has its proof. The proof of 1.1 is just

PROOF: By assumption :1.

Assumption :1 is the first assumption ($r \in \mathbf{Q}$) in the proof of the theorem. (The numbering scheme for assumptions is explained below.) A hierarchical proof must stop somewhere. The general question of where to stop is addressed in Section 4.2. In this proof, we assume the reader understands that the definition of \mathbf{Q} implies that r can be written as the requisite quotient of integers. The proof of 1.2 is the equally simple

¹In their introductory plane geometry course, students in the U.S. are taught to write proofs in a two-column format, the left column containing a sequence of statements and the right column containing their justifications.

PROOF: 1.1 and definition of m and n .

Step 1.3 is proved by a string of equalities, each with a brief justification.

$$\begin{aligned}
 \text{PROOF: } m/n &= \frac{p/\gcd(p, q)}{q/\gcd(p, q)} && [\text{Definition of } m \text{ and } n] \\
 &= p/q && [\text{Simple algebra}] \\
 &= r && [\text{By 1.1}]
 \end{aligned}$$

This type of proof, consisting of a string of equalities, is simple and direct; it works as well for proving any transitive relation, such as $<$, logical equivalence, and implication. It should be used whenever possible.

Step 1.4 has the multistep proof shown in Figure 3, consisting of Steps 1.4.1 through 1.4.3. The “1.4:1” in the proof of Step 1.4.1 denotes assumption 1 (s divides m) in the proof of Step 1.4. The theorem itself is considered to be a step having the null string as its number, which explains why “:1” denotes assumption 1 of the theorem.

1.4. $\gcd(m, n) = 1$

PROOF: By the definition of the gcd, it suffices to:

ASSUME: 1. s divides m

2. s divides n

PROVE: $s = \pm 1$

1.4.1. $s \cdot \gcd(p, q)$ divides p .

PROOF: 1.4:1 and the definition of m .

1.4.2. $s \cdot \gcd(p, q)$ divides q .

PROOF: 1.4:2 and definition of n .

1.4.3. Q.E.D.

PROOF: 1.4.1, 1.4.2, and the definition of gcd.

Figure 3. The proof of Step 1.4.

3. FURTHER DETAILS

3.1. A More Compact Numbering Scheme. The numbering scheme used in the example is fine for short proofs, with few levels of nesting. However, long proofs can have many levels—I often write proofs more than six levels deep. The number 3.1.1.1.1.2 takes a lot of space, and having to distinguish it from 3.1.1.1.2 can soon lead to eye strain.

We eliminate long step numbers by abbreviating 3.1.1.1.2, a five-part step number ending in 2, as $\langle 5 \rangle 2$. Figure 4 shows a fragment of a proof written with the two numbering styles. To understand why abbreviated numbers suffice, consider where Step 3.1.1.1.2 can be used in this proof. The step can be used only after it is proved, but it cannot be used just anywhere after its proof. Step 3.1.1.1.2 cannot be used in the proof of Step 3.1.1.2 because it was proved under the assumption of Step 3.1.1.1, which is different from Step 3.1.1.2’s assumption. The step can be

3.1.1.1. ASSUME: $x \in S$	$\langle 4 \rangle 1$. ASSUME: $x \in S$
PROVE: ...	PROVE: ...
3.1.1.1.1. ...	$\langle 5 \rangle 1$
3.1.1.1.2. ...	$\langle 5 \rangle 2$
3.1.1.1.3. Q.E.D.	$\langle 5 \rangle 3$. Q.E.D.
By 3.1.1.1.1 and assumption 3.1.1.1.	By $\langle 5 \rangle 1$ and assumption $\langle 4 \rangle$.
3.1.1.2. ASSUME: $x \in T$	$\langle 4 \rangle 2$. ASSUME: $x \in T$
PROVE: ...	PROVE: ...
...	...

Figure 4. Part of a proof, with long and abbreviated step numbers.

used only where the assumptions under which it was proved hold, which means that it can be used only within the proof of its parent, Step 3.1.1.1. Step 3.1.1.1.2 is the only one in the proof of its parent with a five-part number ending in 2. Although there can be many proof steps with the same abbreviated number $\langle 5 \rangle 2$, no two of them have the same parent, so at most one of them may be used at any point in the proof. A reference to Step $\langle 5 \rangle 2$ always refers to the most recent step with that number. Part 3 of the statement of Step $\langle 5 \rangle 2$ is numbered $\langle 5 \rangle 2.3$.

References to assumptions can be abbreviated even more. An assumption can be used only in the proof of a step, or the proof of one of its descendants. We let $\langle 5 \rangle$ denote the assumption of the level-five step that is an ancestor of (or is) the current step, and $\langle 5 \rangle :3$ denote the third numbered part of that assumption. Since the statement of the theorem has a zero-part number, its assumption is number $\langle 0 \rangle$.

Figure 5 contains the complete proof of our example, written with the abbreviated numbering scheme.

3.2. Proof by Cases. Proof by cases can be expressed with a *Case* step, where

CASE: Statement of assumption.

is an abbreviation for

ASSUME: Statement of assumption.

PROVE: Q.E.D.

The proof of the final “Q.E.D.” step explains why the cases considered are exhaustive; it is usually simple. Figure 6 illustrates the use of the *Case* construct to structure a proof by induction. Note how Step $\langle 1 \rangle 1$ is used in the proofs of both cases, showing why *Case* steps provide more flexibility than would a strictly hierarchical proof-by-cases construct.

4. HOW GOOD ARE STRUCTURED PROOFS?

4.1. My Experience. Some twenty years ago, I decided to write a proof of the Schroeder-Bernstein theorem for an introductory mathematics class. The simplest proof I could find was in Kelley’s classic general topology text [4, page 28]. Since Kelley was writing for a more sophisticated audience, I had to add a great deal of explanation to his half-page proof. I had written five pages when I realized that Kelley’s proof was wrong. Recently, I wanted to illustrate a lecture on my proof style with a convincing incorrect proof, so I turned to Kelley. I could find nothing wrong with his proof; it seemed obviously correct! Reading and rereading the

Theorem. There does not exist r in \mathbf{Q} such that $r^2 = 2$.

PROOF SKETCH: We assume $r^2 = 2$ for $r \in \mathbf{Q}$ and obtain a contradiction. Writing $r = m/n$, where m and n have no common divisors (Step $\langle 1 \rangle 1$), we deduce from $(m/n)^2 = 2$ and the lemma that both m and n must be divisible by 2 ($\langle 1 \rangle 2$ and $\langle 1 \rangle 3$).

ASSUME: 1. $r \in \mathbf{Q}$
2. $r^2 = 2$.

PROVE: False

$\langle 1 \rangle 1$. Choose m, n in \mathbf{Z} such that

1. $\gcd(m, n) = 1$

2. $r = (m, n)$

$\langle 2 \rangle 1$. Choose p, q in \mathbf{Z} such that $q \neq 0$ and $r = p/q$.

PROOF: By assumption $\langle 0 \rangle 1$.

Let: $m \triangleq p/\gcd(p, q)$

$n \triangleq q/\gcd(p, q)$

$\langle 2 \rangle 2$. $m, n \in \mathbf{Z}$

PROOF: $\langle 2 \rangle 1$ and definition of m and n .

$\langle 2 \rangle 3$. $r = m/n$

PROOF: $m/n = \frac{p/\gcd(p, q)}{q/\gcd(p, q)}$ [Definition of m and n]

$= p/q$ [Simple algebra]

$= r$ [By $\langle 2 \rangle 1$]

$\langle 2 \rangle 4$. $\gcd(m, n) = 1$

PROOF: By the definition of the gcd, it suffices to:

ASSUME: 1. s divides m

2. s divides n

PROVE: $s = \pm 1$

$\langle 3 \rangle 1$. $s \cdot \gcd(p, q)$ divides p .

PROOF: $\langle 2 \rangle 1$ and the definition of m .

$\langle 3 \rangle 2$. $s \cdot \gcd(p, q)$ divides q .

PROOF: $\langle 2 \rangle 2$ and definition of n .

$\langle 3 \rangle 3$. Q.E.D.

PROOF: $\langle 3 \rangle 1$, $\langle 3 \rangle 2$, and the definition of gcd.

$\langle 2 \rangle 5$. Q.E.D.

$\langle 1 \rangle 2$. 2 divides m .

$\langle 2 \rangle 1$. $m^2 = 2n^2$

PROOF: $\langle 1 \rangle 1.1$ implies $(m/n)^2 = 2$.

$\langle 2 \rangle 2$. Q.E.D.

PROOF: By $\langle 2 \rangle 1$ and the lemma.

$\langle 1 \rangle 3$. 2 divides n .

$\langle 2 \rangle 1$. Choose p in \mathbf{Z} such that $m = 2p$.

PROOF: By $\langle 1 \rangle 2$.

$\langle 2 \rangle 2$. $n^2 = 2p^2$

PROOF: $2 = (m/n)^2$ [$\langle 1 \rangle 1.2$ and $\langle 0 \rangle 2$]

$= (2p/n)^2$ [$\langle 2 \rangle 1$]

$= 4p^2/n^2$ [Algebra]

from which the result follows easily by algebra.

$\langle 2 \rangle 3$. Q.E.D.

PROOF: By $\langle 2 \rangle 2$ and the lemma.

$\langle 1 \rangle 4$. Q.E.D.

PROOF: $\langle 1 \rangle 1.1$, $\langle 1 \rangle 2$, $\langle 1 \rangle 3$, and definition of gcd.

Figure 5. A proof of the irrationality of $\sqrt{2}$.

Theorem. All natural numbers are interesting.

ASSUME: n a natural number.

PROVE: n is interesting.

⟨1⟩1. A number is interesting if it is the smallest number not in an interesting set.

PROOF: By definition of interesting.

⟨1⟩2. CASE: $n = 0$

PROOF: By ⟨1⟩1, since 0 is the smallest natural number not in \emptyset .

⟨1⟩3. CASE: 1. $n > 0$

2. $n - 1$ is interesting

PROOF: By ⟨1⟩1, since case assumption ⟨1⟩ implies that

$\{k : k \leq n - 1\}$ is interesting.

⟨1⟩4. Q.E.D.

PROOF: Steps ⟨1⟩2 and ⟨1⟩3, assumption ⟨0⟩, and mathematical induction.

Figure 6. The *Case* construct.

proof convinced me that either my memory had failed, or else I was very stupid twenty years ago. Still, Kelley's proof was short and would serve as a nice example, so I started rewriting it as a structured proof. Within minutes, I rediscovered the error.

My interest in proofs stems from writing correctness proofs of algorithms. These proofs are seldom deep, but usually have considerable detail. Structured proofs provided a way of coping with this detail. The style was first applied to proofs of ordinary theorems in a paper I wrote with Martín Abadi [1]. He had already written conventional proofs—proofs that were good enough to convince us and, presumably, the referees. Rewriting the proofs in a structured style, we discovered that almost every one had serious mistakes, though the theorems were correct. Any hope that incorrect proofs might not lead to incorrect theorems was destroyed in our next collaboration [3]. Time and again, we would make a conjecture and write a proof sketch on the blackboard—a sketch that could easily have been turned into a convincing conventional proof—only to discover, by trying to write a structured proof, that the conjecture was false. Since then, I have never believed a result without a careful, structured proof. My skepticism has helped avoid numerous errors.

I have also found structured proofs very helpful when I need a variant of an existing theorem, perhaps with a slightly weaker hypothesis. In a properly written proof, where every use of an assumption or a proof step is explicit, simple text searching reveals exactly where every hypothesis is used.

4.2. Writing Structured Proofs. A structured proof format by itself will not eliminate errors. Proofs must be written carefully, with enough detail. Most errors come from not carrying out the proof to enough levels. The lowest-level, paragraph-style proofs should be short and completely transparent. One must be a skeptical reader of one's own proofs. My own rule of thumb is to expand the proof until the lowest level statements are obvious, and then continue for one more level. This takes discipline. But, unlike conventional proofs, in which adding more detail can make a proof more confusing, structured proofs accommodate as much detail as desired.

Structured proofs are longer than conventional ones. Although the formatting is partly responsible, structured proofs are longer mainly because they include more

detail. They make it obvious when steps have been forgotten or important details omitted. They make it hard to be sloppy. The assertion “this case is similar to the previous one” is not acceptable; one is forced to find the appropriate general step that makes the proof of both cases easy. Writing a rigorous proof is harder than writing a sloppy one, and lazy writers will find excuses to avoid doing it. A common excuse is that structured proofs are too long. But, shorter proofs are not necessarily better ones; the shortest proof is always “left as an exercise for the reader.”

When journals are distributed electronically, they can include proofs down to the lowest reasonable level; the reader can suppress uninteresting details when viewing the article on the screen or printing it locally. But, for paper journals, extra pages mean killing extra trees. It may be inappropriate for a journal to print a proof with so much detail. I recommend that authors provide two versions of their proofs: a very detailed one for themselves, the referees, and interested colleagues; and a less detailed one for paper publication. It is quite easy to convert a detailed proof into a less detailed one by compressing the lower levels into paragraph-style proofs. Although the reader must fill in the low-level details, such proofs are much better than unstructured ones, in which authors seem to choose randomly which details to supply and which to omit.

4.3. Reading Structured Proofs. So far, readers’ reactions to structured proofs have been mixed. Skeptical readers—ones who check for errors—like these proofs much more than conventional ones. Readers who want to skim the proofs are less happy with the style. Part of the problem is that the length of the proofs and the unfamiliar format are intimidating. The best way to read a structured proof is level by level—first reading the high-level steps $\langle 1 \rangle 1, \langle 1 \rangle 2, \langle 1 \rangle 3, \dots$, then the proofs of those steps, and so on. However, having to skip over the lower-level steps makes reading the high-level ones inconvenient. With hypertext, this is not a problem. With printed text, a layered presentation—the complete higher-level proof followed by the lower-level proofs—may help [2, section B.7 (page 48)].

These structured proofs do not seem ideal for someone who wants to understand the important ideas of a proof without reading any of the details. Satisfying such readers may just require better proof sketches. Or, perhaps a better way of annotating a proof with comments is needed. Hypertext can provide graphical aids for finding one’s way around a proof and highlighting important steps. Maybe such aids can be developed for the printed page.

4.4. The Future. Modern mathematical notation has evolved over hundreds of years. Its proof style is still stuck in the seventeenth century. Mathematicians tend to be conservative, and many are unwilling to consider that there might be a better way of writing proofs. But, I am told that mathematicians are embarrassed to learn that they published incorrect theorems, so they are motivated to avoid errors. I believe they will like structured proofs if they can be persuaded to try them.

Computer scientists are more willing to explore unconventional proof styles. Unfortunately, I have found that few of them care whether they have published incorrect results. They often seem glad that an error was not caught by the referees, since that would have meant one fewer publication. I fear that few computer scientists will be motivated to use a proof style that is likely to reveal their mistakes. Structured proofs are unlikely to be widely used in computer science until publishing incorrect results is considered embarrassing rather than normal.

The proof style described here has been developed over the past several years. I have written many hundreds of pages of structured proofs, mostly of algorithms. I

consider the style to be a great improvement over conventional, unstructured proofs. But, this is not the last word on the subject. I look forward to seeing structured proof styles evolve as mathematicians and computer scientists find better ways to write a proof.

ACKNOWLEDGMENTS. My information about mathematicians' errors and embarrassment comes mainly from George Bergman. The CASE construct and several other details of the proof format were developed in discussions with Urban Engberg and Peter Grønning. Peter Dickman and Lyle Ramshaw found errors in an earlier version.

REFERENCES

1. Martín Abadi and Leslie Lamport. The existence of refinement mappings. *Theoretical Computer Science*, 82(2):253–284, May 1991.
2. Martín Abadi and Leslie Lamport. An old-fashioned recipe for real time. Research Report 91, Digital Equipment Corporation, Systems Research Center, 1992.
3. Martín Abadi and Leslie Lamport. Composing specifications. *ACM Transactions on Programming Languages and Systems*, 15(1):73–132, January 1993.
4. John L. Kelley. *General Topology*. The University Series in Higher Mathematics. D. Van Nostrand Company, Princeton, New Jersey, 1955.
5. Uri Leron. Structuring mathematical proofs. *American Mathematical Monthly*, 90(3):174–185, March 1983.

Digital Equipment Corporation
130 Lytton Avenue
Palo Alto, CA 94301
lamport@src.dec.com

GRAPHING CALCULATOR MATHEMATICS

Today, a *graduate* student in a *second* term of *graduate* complex analysis proudly displayed a program on a graphing calculator to map the unit circle onto an ellipse with a linear fractional transformation. Queried whether “we” had proved early in the first term that linear fractional transformations map generalized circles onto generalized circles, the student then busily checked something on the graphing calculator without first paying any attention to the “proved” theorem. The student’s reaction corroborates rumors of departments about mathematics that promote graphing explorations while forbidding proofs in calculus. In a department of mathematics, however, calls for proofs soon brought the student back to the complexified reality. The incident demonstrates the value of research and assessment of technology in mathematics education and of the resulting graduates’ employability.

Yves Nievergelt
Department of Mathematics
Eastern Washington University
Cheney, WA 99004-2431

Searching for Common Generalizations: The Case of Hyperbolic Functions

Kenneth B. Stolarsky

1. INTRODUCTION. Given two different but related facts, it can be worthwhile to find a single simple statement that contains both of them. For example, Max Planck found after a long deliberate search a law of thermodynamics that included both the Rayleigh Jeans law for long wavelength blackbody radiation and the Wien law for short wavelength blackbody radiation. The significance of this went far beyond having “one less law to remember”; it led to *quantum mechanics*, one of the greatest achievements of twentieth century science. A parallel example in mathematics could have arisen from an attempt to find a good common generalization of the laws

$$\sin(z + 2\pi) = \sin(z) \quad (1.1)$$

and

$$\tanh(z + i\pi) = \tanh(z). \quad (1.2)$$

It does not in fact seem to be the case that any late eighteenth century mathematician deliberately set out upon this task. But it could have led to fame commensurate with Planck’s. With hindsight, define $y(z)$ as the solution to

$$\begin{cases} \frac{dy}{dz} = (1 - y^2)^{1/2} (1 - k^2 y^2)^{1/2} \\ y(0) = 0 \end{cases} \quad (1.3)$$

where $0 \leq k \leq 1$. For $k = 0$ we have $y(z) = \sin z$ and for $k = 1$ we have $y(z) = \tanh z$. For $0 < k < 1$ the function $y(z)$ has a real period $A(k)$ and a purely imaginary period $iB(k)$ where $A(k) \rightarrow 2\pi$ and $B(k) \rightarrow \infty$ as $k \rightarrow 0$ and $A(k) \rightarrow \infty$ and $B(k) \rightarrow \pi$ as $k \rightarrow 1$. This leads to the theory of doubly periodic functions, i.e. *elliptic functions*, one of the greatest achievements of nineteenth century mathematics.

Our theme is the question of what constitutes a good common generalization of previously disparate mathematical facts. In §2 we discuss this in general, with a variety of examples, some of which belong in every mathematician’s repertoire. In the following sections we restrict ourselves to considering various pairs or triplets of facts about *hyperbolic functions*, and create common generalizations. There is no pretense to any finality here—the reader may find other or better generalizations. Perhaps our most curious result is the inequality (6.1). (Since its proof hinges on the “lucky” factorization of Lemma 3 rather than upon any systematic method, it incidentally motivates a question that has been asked before: is there an algorithm for deciding the truth of any “similar” inequality involving hyperbolic functions? (Roughly, can the Tarski algorithm be extended to include real exponentiation?)

There isn't one as yet, but (see §8) mathematical logic has recently shed some light on this in a surprising way.)

We now ask whether "creativity" in any mathematical area can be "mechanized" by the process of selecting pairs of theorems and "systematically" searching for common generalizations. This seems unlikely, especially since there may be (see §2) a vast number of "useless" generalizations. However, there is at least one area of research where simply looking for generalizations (even of *single* facts) has been paying off well for more than a century, the area of "*q*-analogues." Roughly speaking, one tries to replace every integer n by $(q^n - 1)/(q - 1)$ and every derivative $f'(x)$ by $(f(qx) - f(x))/(qx - x)$, where $|q| \leq 1$. Many results of depth and importance have come, e.g., from *q*-analogues of identities between hypergeometric functions. Before going on, we observe that this area provides a very straightforward example of a common generalization. The Taylor expansion

$$f'(x) = \sum_{n=1}^{\infty} \frac{f^{(n)}(0)}{n!} nx^{n-1} \quad (1.4)$$

has the *q*-analogue

$$\frac{f(qx) - f(x)}{qx - x} = \sum_{n=1}^{\infty} \frac{f^{(n)}(0)}{n!} \frac{q^n - 1}{q - 1} x^{n-1} \quad (1.5)$$

which resembles the familiar

$$\frac{f(x) - f(a)}{x - a} = \sum_{n=1}^{\infty} \frac{f^{(n)}(a)}{n!} (x - a)^{n-1}. \quad (1.6)$$

A common generalization of (1.5) and (1.6) is

$$\frac{f(x) - f(qx + (1 - q)a)}{x - (qx + (1 - q)a)} = \sum_{n=1}^{\infty} \frac{f^{(n)}(a)}{n!} \frac{q^n - 1}{q - 1} (x - a)^{n-1}. \quad (1.7)$$

Much of this paper requires no prerequisite beyond calculus. In some places a knowledge of elementary differential equations (and in a parenthetical aside in §3 some rudimentary complex variable theory) would be helpful. Sophisticated readers may view the results of §3 in terms of Riccati equations and those of §5 (see especially (5.5)) in terms of Bernoulli equations. In §7 we use the "big O" notation and the important fact that

$$\lim_{p \rightarrow \infty} \left(1 + \frac{x}{p} + O\left(\frac{1}{p^2}\right) \right)^p = \lim_{p \rightarrow \infty} \left(1 + \frac{x}{p} \right)^p = e^x. \quad (1.8)$$

2. UNIFICATION AND GENERALIZATION. Terms such as "grand unification" are bandied about more by physicists than by mathematicians, so we'll begin with physics. Laws of physics are typically written down as equations, e.g. $E - mc^2 = 0$ and $\nabla \cdot B = 0$. The question has been raised in earnest as to whether all our knowledge of physics can be "unified" and reduced to one single equation. An answer to this is given in volume II of The Feynman Lectures on Physics [F-L-S, pp. 25-10, 25-11]. Let $F_1 = 0, F_2 = 0, \dots, F_n = 0$ be an enumeration of all the equations expressing laws of physics. Then

$$\sum_{i=1}^n F_i^2 = 0 \quad (2.1)$$

is the desired equation! Although initially amusing, I think most regard this as a brazen swindle. It really doesn't tell us anything we didn't already know. [This trick goes back well before Feynman. In fact, it does have some use in mathematical logic. For example, in [S, pp. 2–3] it is used to show that the solution of Diophantine equations of arbitrary degree is no harder than the solution of Diophantine equations of fourth degree. To get a feeling for this, consider the Diophantine equation $x^2 + y^2 = z^5 + 2$. If we set $v = z^2$ and $w = v^2$, so $wz + 2 = x^2 + y^2$, it has become a system of second degree equations. Combine them as above.]

Let's now give illustrations (in mathematics) of good generalizations. It is known that

$$\int_{-\infty}^{\infty} \frac{dx}{x^2 + \pi^2} = 1 \quad (2.2)$$

and

$$\int_{-\infty}^{\infty} \frac{\cos x}{x^2 + 1} dx = \frac{\pi}{e}. \quad (2.3)$$

Both of these are consequences of

$$\int_{-\infty}^{\infty} \frac{\cos \lambda x}{x^2 + a^2} = \frac{\pi}{a} e^{-\lambda a}, \quad a > 0, \lambda \geq 0. \quad (2.4)$$

Compare the *informative* nature of (2.4) with what you get from the scheme of (2.1). In fact (2.4) tells us much more than the two equations it generalizes; for one thing it suggests the Riemann-Lebesgue Lemma [Ap, pp. 469–470], that

$$\lim_{\lambda \rightarrow \infty} \int_{-\infty}^{\infty} g(x) \cos \lambda x dx = 0$$

for well-behaved functions $g(x)$.

Another example is provided by Euler's

$$\sum_{m=-\infty}^{\infty} (-1)^m q^{m(3m-1)/2} = \prod_{n=1}^{\infty} (1 - q^n) \quad (2.5)$$

together with Gauss'

$$\sum_{n=-\infty}^{\infty} (-1)^n q^{n^2} = \prod_{m=1}^{\infty} \frac{1 - q^m}{1 + q^m}. \quad (2.6)$$

Compare what (2.1) produces with Jacobi's elegant [An, p. 21] formula

$$\sum_{n=-\infty}^{\infty} z^n q^{n^2} = \prod_{n=0}^{\infty} (1 - q^{2n+2})(1 + zq^{2n+1})(1 + z^{-1}q^{2n+1}). \quad (2.7)$$

There is, however, a bit to be said in favor of (2.1). It doesn't introduce additional parameters, and one can recover the special cases rather easily. The additional parameters could of course be regarded as a bonus, but one does need a bit of cleverness to deduce (2.5) and (2.6) from (2.7). In fact, given propositions P and Q, it is not always easy to decide whether or not P is a generalization of Q. This difficulty will arise again in connection with the result of §6. For some further discussion of this see [B, p. 34].

We now turn to generalizations of inequalities. For example, can

$$\frac{2ab}{a+b} \leq \sqrt{ab} \quad a, b \geq 0 \quad (2.8)$$

and

$$\frac{a+b}{2} \leq \left(\frac{a^3+b^3}{2} \right)^{1/3} \quad a, b \geq 0 \quad (2.9)$$

be unified? From [B-B, pp. 17-18] we find that

$$\left(\frac{a^s+b^s}{2} \right)^{1/s} \leq \left(\frac{a^t+b^t}{2} \right)^{1/t} \quad a, b \geq 0, -\infty < s \leq t < \infty. \quad (2.10)$$

By taking $s = 1$ and $t = 3$ we recover (2.9), but it is not obvious that (2.8) is a special case.

One needs to take $s = -1$ and consider the *limiting case* $t \rightarrow 0$ (Hint: take logarithms and apply L'Hopital's rule).

Is it true that any system of inequalities has a common generalization? Say $F_1(a, b) \geq 0, F_2(a, b) \geq 0, \dots, F_N(a, b) \geq 0$ for (a, b) in some region R . Let

$$L_i(t) = \frac{(t-1)^2(t-2)^2 \cdots (t-N)^2}{(t-i)^2}.$$

Then

$$\sum_{i=1}^N L_i(t) F_i(a, b) \geq 0, \quad (a, b) \in R, t \text{ real} \quad (2.11)$$

is a common generalization. It is clearly true, and for $t = j$ it reduces to the j th inequality $F_j(a, b) \geq 0$. However, this has all the faults of (2.1), and the additional "fault" of introducing another parameter.

3. TWO INTEGRALS. Let $\theta \geq 0$. Then

$$\int_0^\infty \left(\frac{1}{x + \coth \theta} \right)^2 dx = \tanh \theta \quad (3.1)$$

and

$$\int_0^\infty \tanh^2(x + \theta) dx \text{ is divergent.} \quad (3.2)$$

These are quite simple, but it is not clear that they belong together. Is there a good common generalization? A small clue is provided by the theorem below that describes the solutions of a certain type of differential equation.

Theorem. For t real, let $f(t)$ and $g(t)$ be functions that are nonnegative and continuously differentiable with $f(0) = 0, f'(t) > 0$ for $t > 0, g(t) > 0$ for $t \geq 0$, and

$$\lim_{x \rightarrow \infty} g(x) = 0. \quad (3.3)$$

Let $u_0 \geq 0$. If $u = u(x)$ satisfies

$$\begin{cases} \frac{du}{dx} = g(x) - f(u) \\ u(0) = u_0 \end{cases} \quad (3.4)$$

for $0 \leq x < \infty$, then $u(x) \rightarrow 0$ and $u'(x) \rightarrow 0$ as $x \rightarrow \infty$.

Proof: If $u_0 = 0$ then $\left. \frac{du}{dx} \right|_{x=0} = g(0) - f(0) > 0$ and $u(x)$ is positive in some deleted right neighborhood of 0; if $u_0 > 0$ this is true by continuity. If $u(x) = 0$ for some $x > 0$ there is a smallest such x , say x_0 . We then have that $u'(x_0) = g(x_0) - f(0) > 0$ and hence $u(x)$ is negative in some deleted left neighborhood of x_0 . The Intermediate Value Theorem now shows that $u(x_1) = 0$ for some x_1 with $0 < x_1 < x_0$, a contradiction. Hence $u(x) > 0$ for $x \geq 0$. Now let $\epsilon > 0$. If $u(x) \geq \epsilon$ for all sufficiently large x , then

$$\begin{aligned} u'(x) = g(x) - f(u(x)) &\leq \left(g(x) - \frac{f(\epsilon)}{2} \right) - \frac{f(\epsilon)}{2} \\ &\leq -f(\epsilon)/2 \end{aligned}$$

for all sufficiently large x , since $f(t)$ is increasing. This implies $u(x)$ will become negative, a contradiction. Hence there is a sequence of real numbers $x_1 < x_2 < x_3 < \dots$ with $x_n \rightarrow \infty$ such that $u(x_n) < \epsilon$. Choose n so large that $g(x) < \frac{1}{2}f(2\epsilon)$ for $x \geq x_n$, and let x^* be the smallest x such that $x \geq x_n$ and $u(x) \geq 2\epsilon$. Clearly x^* is strictly to the right of x_n . However

$$u'(x^*) < \frac{1}{2}f(2\epsilon) - f(2\epsilon) < 0,$$

so $u(x)$ is greater than 2ϵ in some deleted left neighborhood of x^* , a contradiction. Hence $u(x) < 2\epsilon$ for all $x \geq x_n$ and since $\epsilon > 0$ was arbitrary, this says that $u(x) \rightarrow 0$ as $x \rightarrow \infty$. It now follows from the differential equation that $u'(x) \rightarrow 0$ as $x \rightarrow \infty$.

We now consider the function $u(x)$ defined by

$$\begin{cases} u'(x) = e^{-ax} - u^2(x) \\ u(0) = \tanh \theta \end{cases} \quad (3.5)$$

where $\theta \geq 0$ and $0 \leq a \leq \infty$. [Here are some comments for readers concerned with technical rigor. To see that a function exists for *all* $x \geq 0$ and is unique, observe that the coefficients of the *linear* second order differential equation

$$v'' - e^{-ax}v = 0$$

have no finite singularities, so there is a general solution $v = c_1v_1(x) + c_2v_2(x)$ that is analytic for all finite x with the Wronskian of v_1 and v_2 never vanishing. This means that some such v can be found that satisfies $v'(0)/v(0) = \tanh \theta$, and a straightforward calculation shows that $u(x) = v'(x)/v(x)$ satisfies the differential equation. Moreover, $v'(x)/v(x)$ can have as singularities only simple poles, and if some real $x_0 > 0$ is a pole of $u(x)$, then $u(x)$ (which is always positive) must approach $+\infty$ as $x \rightarrow x_0$ from the left. However, $u'(x)$ is then always negative in a deleted left neighborhood of x_0 , a contradiction. Hence $u(x)$ exists for all $x \geq 0$; its uniqueness is an easy consequence of the standard uniqueness theorem.]

By the previous Theorem, $\lim_{x \rightarrow \infty} u(x) = 0$ so

$$\begin{aligned} \int_0^\infty u^2(x) dx &= \int_0^\infty e^{-ax} dx - \int_0^\infty u'(x) dx \\ &= \frac{1}{a} - \left[\lim_{b \rightarrow \infty} u(b) - u(0) \right] \\ &= \frac{1}{a} + \tanh \theta. \end{aligned} \quad (3.6)$$

We recover (3.1) by letting $a \rightarrow \infty$ and (3.2) by letting $a \rightarrow 0$.

4. A NONLINEAR GENERALIZATION OF HYPERBOLIC FUNCTIONS. Let $x \geq 0$. Consider the following facts. If

$$y_1(x) = \sinh x \text{ and } y_2(x) = \cosh x \text{ then } y_2(x) - y_1(x) \rightarrow 0 \text{ as } x \rightarrow \infty. \quad (4.1)$$

If

$$y_1(x) = x \text{ and } y_2(x) = x + 1 \text{ then } y_2(x) - y_1(x) = 1 \text{ for all } x. \quad (4.2)$$

If

$$y_1(x) = \sqrt{x+1} \text{ and } y_2(x) = 2\sqrt{x+1} \text{ then } y_2(x) - y_1(x) \rightarrow \infty \text{ as } x \rightarrow \infty. \quad (4.3)$$

One might say that (4.1) is trivial while (4.2) and (4.3) insult your intelligence, but what is a good common generalization?

Theorem. If $y_1 = y_1(x)$ and $y_2 = y_2(x)$ are functions satisfying $0 \leq y_1(0) < y_2(0)$ and the differential equations

$$\frac{dy_1}{dx} = Cy_2^\alpha, \quad \frac{dy_2}{dx} = Cy_1^\alpha \quad (4.4)$$

where $C > 0$, then for some constant c_0

$$y_2(x) - y_1(x) \rightarrow \begin{cases} 0 & \alpha > 0 \\ c_0 & \alpha = 0 \\ \infty & \alpha < 0 \end{cases} \quad (4.5)$$

as x increases without limit in the (possibly infinite) domain of definition of $y_1(x)$ and $y_2(x)$.

Before proving this we observe that the case $\alpha = 0$ is trivial, while $\alpha = 1$ is straightforward since one then obtains simple *linear* differential equations for y_1 and y_2 . In particular, for $C = 1$ and $0 = y_1(0) < y_2(0) = 1$, we have the familiar equations

$$\frac{d}{dx} \sinh x = \cosh x, \quad \frac{d}{dx} \cosh x = \sinh x. \quad (4.6)$$

For $\alpha > 1$ it is not hard to show that there is a finite x_0 such that

$$\lim_{x \rightarrow x_0} y_1(x) = \lim_{x \rightarrow x_0} y_2(x) = \infty.$$

The theorem is at any rate plausible, since for $\alpha > 0$ the smaller function always has the larger derivative while the opposite is true for $\alpha < 0$. However, the only case other than $\alpha = 0, 1$ in which the author can explicitly solve for y_1 and y_2 (and

thus establish the result directly) is $\alpha = -1$; we leave this as a simple exercise that “yields” (4.3). The idea of our proof is to exploit a relationship that generalizes the familiar $\cosh^2 x - \sinh^2 x = 1$ relation that follows from (4.6) by differentiation.

Proof: Clearly y_1 and y_2 are non-decreasing. When $\alpha > 0$ we have from $y_2(0) > 0$ that $y_1(x) \rightarrow \infty$ and hence that $y_2(x) \rightarrow \infty$. Say $\alpha < 0$. If y_1, y_2 are bounded, their derivatives will both have positive lower bounds, a contradiction. Hence one of them will approach ∞ as x increases. Now observe that for $\alpha \neq -1$,

$$y_2(x)^{\alpha+1} - y_1(x)^{\alpha+1} = y_2(0)^{\alpha+1} - y_1(0)^{\alpha+1} = K \neq 0 \quad (4.7)$$

for some constant K : just differentiate! If $\alpha > 0$ then $K > 0$ and the Mean Value Theorem yields

$$K = (y_2 - y_1)(\alpha + 1)\xi^\alpha$$

where ξ is between $y_1(x)$ and $y_2(x)$, so

$$y_2 - y_1 \leq K/(\alpha + 1)y_1(x) \rightarrow 0$$

as x increases. For $-1 < \alpha < 0$ set $z_2(x) = y_2^{1+\alpha}$ and $z_1(x) = y_1^{1+\alpha}$. Then again $K > 0$ and

$$\begin{aligned} y_2(x) - y_1(x) &= z_2^{1/(1+\alpha)} - z_1^{1/(1+\alpha)} \\ &= (z_2 - z_1) \frac{1}{1 + \alpha} \xi^{(-\alpha)/(1+\alpha)} \\ &\geq \frac{K}{1 + \alpha} z_1(x)^{-\alpha/(\alpha+1)} = \frac{K}{1 + \alpha} y_1(x)^{(-\alpha)}. \end{aligned}$$

In this case we have $y_2(x) \rightarrow \infty$, but this clearly implies that $y_1(x) \rightarrow \infty$ and the result follows. Finally, say $\alpha < -1$. Then

$$\frac{1}{y_2^{-\alpha-1}} - \frac{1}{y_1^{-\alpha-1}} = K < 0 \quad (4.8)$$

so both of y_1 and y_2 cannot approach infinity as x increases. However, we know that one of them must approach ∞ . Clearly it is y_2 , and the result follows.

We remark that if $\alpha > -1$ and $y_1(0) = 0$, the ratio

$$t(x) = t(x; \alpha, C) = y_1(x)/y_2(x)$$

will resemble the hyperbolic tangent in several ways. From (4.7) it is clear that $t(x) \rightarrow 1$ as x increases, and

$$D^2(y_1/y_2) = D(CK/y_2^2) = \frac{-2C^2Ky_1^\alpha}{y_2^3} < 0$$

so the function is concave downwards. Also, the inequality $\tanh x \leq x$ has the analogue

$$\frac{y_1}{y_2} \leq \frac{CK}{y_2^2(0)}x.$$

Finally, $t(x)$ solves the differential equation

$$y' = CK^{(\alpha-1/\alpha+1)}(1 - y^{\alpha+1})^{2/(\alpha+1)}$$

Let us say that we interpret this for $\alpha = -1$ as the statement “ y is constant.” Then we have the curious fact that for $\alpha = -5, -3, -2, -1, 0, 1, 3$ the equation

can be integrated in terms of familiar functions (indeed *elementary* functions aside from -5 and 3 which require elliptic functions). Perhaps the “transelliptic” integrals that $\alpha = 2$ and $\alpha = -4$ lead us to are worthy of special attention?

5. SOME HYPERBOLIC INEQUALITIES. The identities

$$\sinh 2u = 2 \sinh u \cosh u \quad (5.1)$$

and

$$\cosh^2 x - \sinh^2 x = 1 \quad (5.2)$$

suggest (in very different ways) certain inequalities. First, what happens if each 2 is replaced by x in (5.1)? We shall see that if $0 \leq x \leq 2$ and $u \geq 0$, the left side is then at most as large as the right side. Next, observe that integration of $1 \leq \cosh x$ yields $x \leq \sinh x$ and hence

$$x^2(\cosh^2 x - \sinh^2 x) \leq \sinh^2 x. \quad (5.3)$$

A slight rearrangement here gives

$$a(x) := \frac{x}{\sqrt{1+x^2}} \leq \tanh x. \quad (5.4)$$

This is fairly tight; numerical calculations indicate that the difference of the two sides is at most .073688... (near $x = 1.6219...$). Both of the above functions are concave downwards, behave like x as $x \rightarrow 0$, and both approach 1 as $x \rightarrow \infty$. Hence if C is any constant and $f(x)$ denotes either of them, we expect

$$\int_0^x f(t) dt \sim x \sim x - C \sim \frac{x}{f(x)} - C$$

as $x \rightarrow \infty$. In fact we have the identity

$$\int_0^x a(t) dt = \frac{x}{a(x)} - 1. \quad (5.5)$$

Now, what happens if in (5.5) we replace each $a(\cdot)$ by $\tanh(\cdot)$? We shall see that here it is the right side that becomes smaller. Thus our program now is to find a common generalization of

$$\sinh xu \leq x \cosh u \sinh u, \quad 0 \leq x \leq 2, 0 \leq u \quad (5.6)$$

and

$$0 \leq x \coth x - 1 \leq \log \cosh x, \quad 0 \leq x. \quad (5.7)$$

First, however, let's show they are true. Inequality (5.6) may be rewritten as

$$f(u) = \sinh xu \leq x \frac{\sinh 2u}{2} = g(u).$$

Here $f(0) = g(0)$ and $f'(u) \leq g'(u)$ is

$$x \cosh xu \leq x \cosh 2u,$$

and (5.6) follows. From the obvious inequality $\sinh x \leq \cosh x$ we deduce

$$\tanh x \leq 1 \leq \coth x. \quad (5.8)$$

A differentiation argument similar to the above (and even easier) shows that

$$\tanh x \leq x \leq \sinh x, \quad x \geq 0. \quad (5.9)$$

Now the left side of (5.7) is nothing more than the left side of (5.9). The right side of (5.7) is more subtle; it will be established by an appropriate introduction of a $2 \sinh x \cosh x$ term. The attempt to prove it by differentiation leads to the perhaps unfamiliar inequality

$$\coth x - \frac{x}{\sinh^2 x} \leq \tanh x; \quad (5.10)$$

is this true? The left side of (5.9) yields

$$2 \cosh x \sinh x - 2x \leq 2 \cosh x \sinh x - 2 \tanh x$$

and the right side of the above is $(\cosh^2 x - \sinh^2 x = 1)$

$$2(\cosh^2 x - 1)\tanh x = \frac{2 \sinh^3 x}{\cosh x}.$$

Cancelling twos gives

$$\cosh x \sinh x - x \leq \tanh x \sinh^2 x$$

and division by $\sinh^2 x$ yields (5.10) and hence the result.

6. THE COMMON GENERALIZATION. Here we establish the

Theorem. Let $0 \leq s \leq 1$, $0 \leq u$, and $1 \leq p$. Then

$$\left(1 + \frac{s}{p}\right)^p \leq \left(\frac{\sinh(p+s)u}{\sinh pu}\right)^p \leq \left(1 + \frac{s}{p}\right)^p \cosh pu. \quad (6.1)$$

Perhaps this does not seem to bear any relation to (5.6) and (5.7); we'll address this concern later. Since $\lim_{x \rightarrow 0} \sinh x/x = 1$, i.e., $\sinh x \sim x$ as $x \rightarrow 0$, the theorem is certainly true in the limiting case $u \rightarrow 0$. For $p \rightarrow \infty$ it implies that the middle term tends to e^s provided that u goes to zero faster than $1/p$.

The proof requires some preliminary results; the essential key to the proof is Lemma 3.

Lemma 1. Let $x, u \geq 0$. Then

$$f(x) = \frac{\sinh x}{x} \text{ and } g(x) = x \coth xu \quad (6.2)$$

are increasing in x .

Proof: We have

$$f'(x) = \frac{(\cosh x)(x - \tanh x)}{x^2} \geq 0 \text{ and } g'(x) = \frac{(\sinh 2xu) - 2xu}{2 \sinh^2 xu} \geq 0.$$

Lemma 2. Let $u \geq 0$, $p \geq 1$. Then

$$a(u) = \tanh pu \leq p \tanh u = b(u). \quad (6.3)$$

Proof: Both sides are 0 when $u = 0$ and $a'(u) \leq b'(u)$ is

$$\frac{p}{\cosh^2 pu} \leq \frac{p}{\cosh^2 u}.$$

Lemma 3. For $u \geq 0$ and $p \geq 1$ we have

$$(p+1)\coth(p+1)u - p \coth pu \leq \tanh pu \quad (6.4)$$

Proof: The above may be written, in an obvious notation, as

$$\begin{aligned} \frac{C}{D} &= \frac{1 + \tanh pu \tanh u}{\tanh pu + \tanh u} \\ &= \coth(p+1)u \leq \frac{p^{-1/2} \tanh pu + p^{1/2} \coth pu}{p^{-1/2} + p^{1/2}} = \frac{A}{B}. \end{aligned}$$

Thus the assertion is that $\coth(p+1)u$ is at most a certain weighted average of $\tanh pu$ and $\coth pu$. The author does not find this at all obvious, but it happens to be true that

$$\frac{AD - BC}{BD} = \frac{1}{BD} p^{-1/2} (1 - \tanh^2 pu) \left(\frac{p \tanh u}{\tanh pu} - 1 \right) \geq 0;$$

recall (5.8) and Lemma 2!

We now prove the Theorem by logarithmic differentiation. It is also helpful to make the replacement

$$\left(1 + \frac{s}{p}\right)^p \rightarrow (p+s)^p/p^p$$

and associate $(p+s)^p$ and p^p with $(\sinh(p+s)u)^p$ and $(\sinh pu)^p$ respectively. The left inequality now follows easily from the property of $f(x)$ given in Lemma 1. For the right side it suffices to show that

$$\frac{d}{du} \left(\log \frac{\sinh(p+s)u}{p+s} - \log \frac{\sinh pu}{p} \right) \leq \frac{d}{du} \log \cosh pu,$$

i.e. that

$$(p+s)\coth(p+s)u - p \coth pu \leq \tanh pu.$$

By the property of $g(x)$ given in Lemma 1, this would follow from the most extreme case in which $s = 1$. But this is simply Lemma 3, and we are done!

7. IS IT A GENERALIZATION?. If we set $p = 1$ the right inequality of (6.1) becomes

$$\sinh(1+s)u \leq (1+s)\sinh u \cosh u,$$

and upon setting $x = 1+s$ we obtain (5.6). To extract (5.7) let $u = x/p$ and let $p \rightarrow \infty$. Thus

$$e^s \leq \lim \left(\cosh \frac{sx}{p} + \sinh \frac{sx}{p} \coth x \right)^p \leq e^s \cosh x.$$

Since

$$\cosh x = 1 + O(x^2) \text{ and } \sinh x = x + O(x^3)$$

the expression raised to the p th power above is

$$\left(1 + \frac{sx \coth x}{p} + o\left(\frac{1}{p^2}\right) \right)^p \rightarrow e^{sx \coth x}$$

as $p \rightarrow \infty$, so

$$1 \leq e^{sx \coth x - s} \leq \cosh x;$$

now take logarithms.

8. CONCLUDING REMARKS. The proof of the Theorem hinged upon some rather good luck at one point (Lemma 3). Is it possible that there is an algorithm that will mechanically prove or disprove any such inequality? No one knows, but a wonderful result has recently been proved by A. J. Macintyre and A. J. Wilkie [M-W]. If Schanuel's conjecture regarding transcendental numbers is true, then any inequality

$$F(x_1, \dots, x_n) \geq 0,$$

where the function F is formed by any finite number of rational operations and real exponentiations, is decidable true or false! Schanuel's conjecture is that the transcendence degree of

$$K = \mathbb{Q}(\alpha_1, \dots, \alpha_n, e^{\alpha_1}, \dots, e^{\alpha_n}),$$

which we denote by $\text{tdeg } K$, is at least n when $\alpha_1, \dots, \alpha_n$ are linearly independent over the rational field \mathbb{Q} . (In particular $\text{tdeg } \mathbb{Q}(\sqrt{2}, e^{\sqrt{2}}) \geq 1$, so one of $\sqrt{2}$ and $e^{\sqrt{2}}$ is not algebraic: this particular case of the conjecture, and a few others, are already known). Thus every proof of a hyperbolic inequality can be regarded as evidence (perhaps infinitesimal evidence) in favor of Schanuel's conjecture.

We end by asking if we can decide whether two hyperbolic inequalities, say $H_1 \geq 0$ and $H_2 \geq 0$, have a "good" common generalization. This seems hard even to formulate; for one thing neither "good" nor "generalization" have been formally defined. Moreover, if $H_3 \geq 0$ is any further "unrelated" inequality, then it could be claimed that the scheme of (2.11) applied to H_1 , H_2 and H_3 yields a truly informative common generalization of H_1 and H_2 . The right way of posing the problem remains to be found.

REFERENCES

- [An] G. E. Andrews, *The Theory of Partitions*, Addison-Wesley, Reading, 1976.
- [B] R. Bellman, *A Brief Introduction to Theta Functions*, Holt, Reinhart and Winston, New York, 1961.
- [B-B] E. F. Beckenbach and R. Bellman, *Inequalities*, Springer-Verlag, Berlin, 1983.
- [F-L-S] R. P. Feynman, R. B. Leighton, and Matthew Sands, *The Feynman Lectures on Physics*, vol. II, Addison-Wesley, Reading, c. 1963.
- [M] D. S. Mitrinović, *Analytic Inequalities*, Springer-Verlag, Berlin, 1970.
- [M-W] A. J. Macintyre and A. J. Wilkie, *On the decidability of the real exponential field*, preprint.
- [S] T. Skolem, *Diophantische Gleichungen*, Ergebnisse der Mathematik 5 (1938), Springer-Verlag, Berlin.

Department of Mathematics
University of Illinois
1409 W. Green Street
Urbana, IL 61801

Answer to Picture Puzzle (p. 594)

Wendy and Alex Robertson—not the only man and wife team of mathematicians on record, but one of the few such.

Transforming n -gons by Folding the Plane

P. Sabinin and M. G. Stone

1. INTRODUCTION. A number of authors have investigated labelled polygonal rings in the plane, and the reconstruction of these rings from distorted images. This reconstruction problem originates in Biology, where the accepted model views chromosomes as the vertices of a polygonal ring ([1], [2]). Invasive experimental techniques appear to distort this ring structure. Folding is one possible form of distortion, and we examine the effect of folding on the distribution of points in a (regular) polygonal ring. Folding the plane transforms a regular n -gon into a set of n (not necessarily distinct) images. We prove, for example, that every set of three points in the plane is the image of a suitable equilateral triangle under a single fold. Every four points can be obtained from a suitable square by a sequence of at most three folds. Some experimentation leads to the natural conjecture that each arbitrary n -point configuration is the image of a suitable *regular* n -gon under finitely many folds. We prove that this is true and establish bounds for the required number of folds.

2. THREE POINTS. We consider transformations of the plane obtained by folding the plane along a line. A *fold* along a line L leaves each point fixed in one half plane determined by L , and reflects each point in the other half plane into its mirror image in L . If a point x is transformed under a fold, we will denote the image of x by x' , and rely upon context to identify the precise nature of the fold itself. Line segments connecting points a and b will be denoted by \overline{ab} , and the length of \overline{ab} is $|\overline{ab}|$. We indicate that x belongs to the line segment \overline{ab} by writing $x \in \overline{ab}$.

We are interested first in the locus of points which are accessible from some given point c by a single fold. Without further restrictions, clearly every point d is accessible from c by a single fold along L , the perpendicular bisector of \overline{cd} . If we further require that another point a is fixed under the fold, then exactly those points d which are interior to (including those on the boundary of) the circle with center a and radius $r = |\overline{ac}|$ are so accessible. If a certain pair of points a and b are both to remain individually fixed under the fold, then the points accessible from c by such a fold are precisely those which lie in the "lens" determined by c between a and b , as indicated in Figure 1. Folding c to d , for d in the (shaded) lens determined by c between a and b , leaves both a and b fixed. This is easily seen since L , the perpendicular bisector of \overline{cd} , fails to meet \overline{ab} ; indeed \overline{cd} is part of a chord on the circles which determine the lens, and the bisector of such a chord passes through the endpoint of \overline{ab} which is the center for the circle in question.

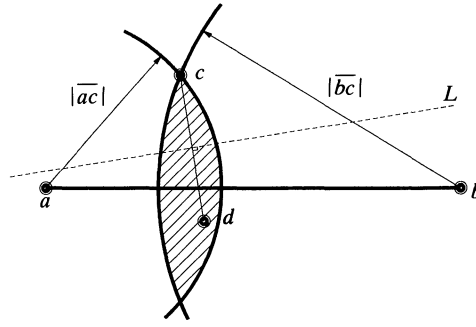


Figure 1. The “lens” determined by c between a and b .

More generally:

Lemma 1. *The points accessible from c by a single fold which leaves a_1, \dots, a_n fixed are exactly those points interior to, or on the boundary of, the intersection of the circles C_i through c with centers at a_i , $i = 1, \dots, n$.*

Proof: The Lemma is obvious for $n = 1, 2$. Consider $k + 1$ points where $k \geq 2$. Let $d \in \bigcap_{i=1}^{k+1} C_i$ and fold c to d along L , the perpendicular bisector of \overline{cd} . Note for each $j = 1, 2, \dots, n$ that $d \in C_j$ so the perpendicular bisector of \overline{cd} passes through the line segment $\overline{ca_j}$ (including the endpoints), hence the fold along L leaves each point a_j fixed, $j = 1, \dots, n$. \square

It is also useful to observe:

Lemma 2. *For any fold along a line L we have $|\overline{ab}| \geq |\overline{a'b'}|$, that is the distance between any two points can only decrease under folding.*

Proof: This is an easy consequence of the triangle inequality. \square

Theorem 1. *Given any three points in the plane, a, b, c , there is an equilateral triangle T with vertices x, y, z , for which a, b , and c are the images of x, y , and z under a single fold.*

Proof: Label the points so that $|\overline{ab}|$ is the largest distance between any two of the three given points. The remaining point c then lies inside one half of the lens which is the intersection of two circles of radius $|\overline{ab}|$ having centers at a and b respectively. (See Figure 2.) Let T be either of the two equilateral triangles determined by the segment \overline{ab} , with vertices $x = a, y = b$, and z . Notice that a fold along the line L which is the perpendicular bisector of \overline{cz} takes z to c and leaves both a and b fixed by Lemma 1. \square

3. FOUR POINTS. We conjecture that every four points are the images of the vertices of a suitable square under at most two folds. We prove two weaker results:

Theorem 2. *Given any four points in the plane a, b, c , and d , there is some square S with vertices x, y, z , and w for which a, b, c , and d are the images of x, y, z , and w under a sequence of at most three folds.*

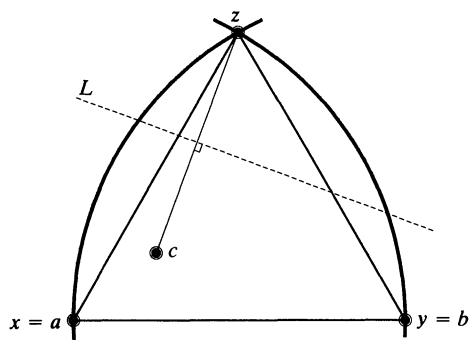


Figure 2. Points accessible from an Equilateral Triangle.

Proof: There are three cases according to the configuration of the four given points: the points may be the vertices of a rectangle, or be collinear, or neither of these. If the points form a rectangle, obviously one fold will suffice. If the four points are collinear, we construct a square whose diagonal contains a , b , c , and d and has one of the two extreme points, say a , as a vertex. Choose the square sufficiently large that all four points lie in one half of the diagonal (see Figure 3a). By Lemma 1, a fold of the opposite diagonal corner z to the remaining extreme

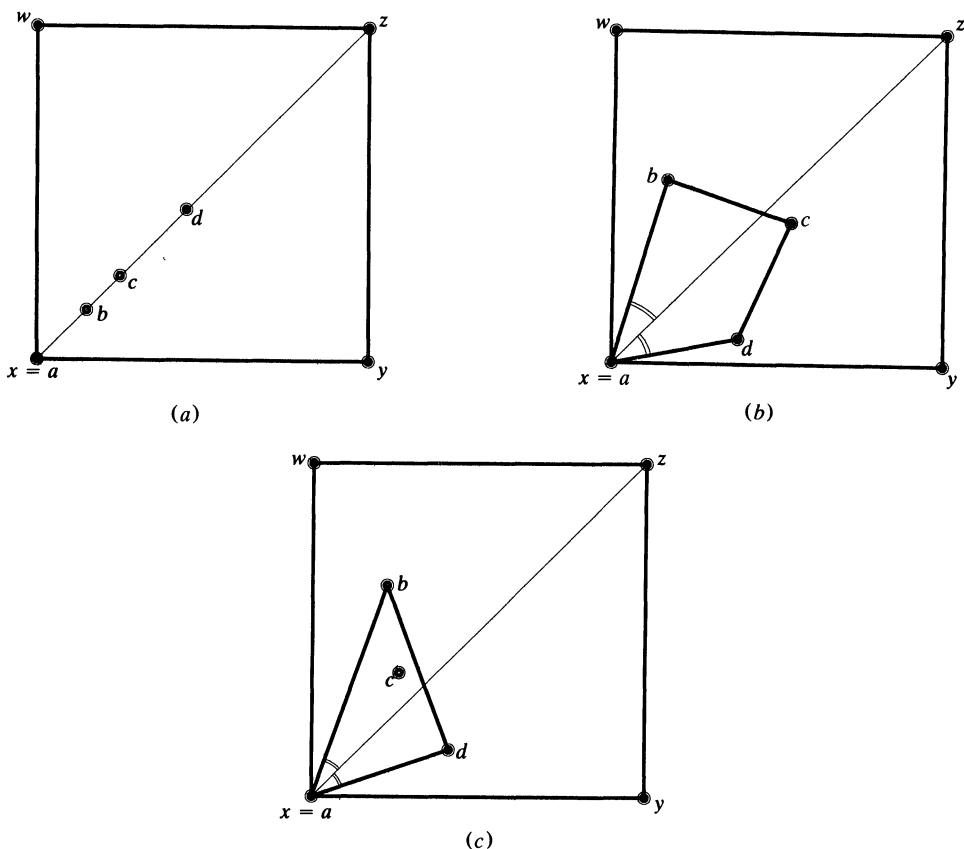


Figure 3. (a) Four Collinear Points. (b) A Quadrilateral. (c) One Interior Point

point d leaves w and y fixed. Each of w and y may then be folded to b and c to complete the sequence.

Finally, in the remaining case, we may assume there are four non-collinear points, and that one interior angle of their convex hull is less than 90° . Label these points a, b, c , and d so that a is the vertex at the apex of a minimal interior angle. Insure also that the point diagonally opposite (if the convex hull is a quadrilateral) or interior (if the hull is a triangle) is labelled c . We then construct a square with vertices x, y, z, w with $a = x$, such that the diagonal of the square bisects the interior angle at a (see Figure 3b and 3c).

By choosing the length of the side of the square to be sufficiently large, we can also guarantee the following conditions:

1. Each of the perpendicular bisectors of \overline{ad} and \overline{ab} intersects each of the sides of the square xy and xw .
2. In case $abcd$ is a quadrilateral, the perpendicular bisector of the diagonal \overline{ac} also intersects both xy and xw .
3. The entire convex region $abcd$ is contained within the lower half of the lens determined by z , between y and w .

Now the square $xyzw$ may be folded to produce $abcd$ as follows:

First, fold z to the point closest to the diagonal xz .

Second, fold y to the remaining point closest to y .

Finally, fold w to the only remaining point.

Lemma 1 guarantees, using conditions 1, 2, and 3, that each of these folds leaves a, b, c, d fixed. Thus each successive fold fails to interfere with the work of the previous folds, and $xyzw$ are taken to a, b, c, d by a sequence of three folds. \square

Theorem 3. *Any four collinear points are the images of the vertices of a suitable square under, at most, two folds.*

Proof: Let the four collinear points be labelled a, b, c, d . By a suitable change in the labelling, we may assume a and d are the extreme points; $b, c \in \overline{ad}$; and the points are labelled so that $|\overline{ac}| < |\overline{ab}|$ and $|\overline{ac}| \leq |\overline{bd}|$. That is, we assume that $|\overline{ac}|$ is the shortest distance between an interior point and the extreme points a and d (the length of $|\overline{ac}|$ is exaggerated in Figure 4 to display, more clearly, the relationship between other points).

We introduce coordinates, as shown in Figure 4, with $|\overline{ad}| = 1$ representing unit length. The idea is simple: we are to fold the point labelled q at $(1, 0)$ to c leaving a and d fixed. Then fold the image p' of the point $p = (0, 0)$ to b with a second

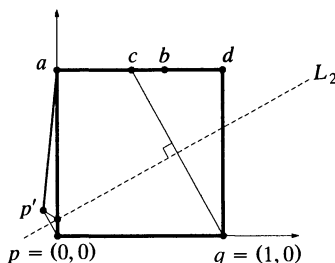


Figure 4. Four collinear points, appropriately labelled.

fold which leaves a , c , and d fixed as well. The first part of this is easy: fold q to c by folding along L_2 the perpendicular bisector of the line segment qc ; c lies within the lens determined by q between a and d , and c as well as a and d remain fixed under this fold.

To assist in describing the image p' of $p = (0, 0)$ under the first fold, we note that the line L_1 through $c = (k, 1)$ and $q = (1, 0)$ has parametric form: $(x, y) = (1 + t, -1/(1 - k)t)$ and the line L_2 is given by $(x, y) = (k + 1/2 + t, \frac{1}{2} + (1 - k)t)$. We deduce that the y intercept for L_2 is $k^2/2$, and we shall use this to help estimate $|\overline{ap'}|$. The real problem here is to see that $|\overline{ap'}|$ is large enough for p' to reach b in the next fold, while leaving a fixed.

Next, we note that $k = |\overline{ac}| < \frac{1}{2}$; this follows from $|\overline{ac}| \leq |\overline{bd}|$ since $|\overline{ac}| \geq \frac{1}{2}$ yields $|\overline{bd}| \geq \frac{1}{2}$ and $|\overline{ad}| = |\overline{ac}| + |\overline{bc}| + |\overline{bd}| > 1$. Thus $k^2 < k/2$. Also, $|\overline{ap'}| + (k^2/2) + (k^2/2) > |\overline{ap}| = |\overline{ad}|$, since the length of the path: $a \rightarrow p' \rightarrow (0, k^2/2) \rightarrow p$ is clearly longer than the direct path $|\overline{ap}|$. We conclude that $|\overline{ap'}| > |\overline{ad}| - k^2 > |\overline{ad}| - k/2$. Moreover, $|\overline{ad}| - (k/2) \geq |\overline{ab}|$ since $|\overline{bd}| \leq |\overline{ac}|$. Finally, then, $|\overline{ap'}| > |\overline{ab}|$, and b is accessible from p' by a fold which leaves a itself fixed, since b lies within the lens determined by p' between a and d . This same fold leaves both c and d fixed as well. Thus the square is transformed, by a sequence of two folds, into the four given collinear points. \square

4. BOUNDS FOR $n \geq 5$. Theorems 1 and 2 lead to the natural conjecture that every n -point set is the image of the vertices of some regular n -gon under finitely many folds. We first prove this directly for $n = 5$.

Theorem 4. *Every five points are the images of the vertices of a suitable regular pentagon under at most five folds.*

Proof: Observe that the lens determined by each vertex between the two adjacent vertices represents the set of points accessible from the given vertex by a fold which (by Lemma 1) leaves all four remaining vertices fixed. It is only necessary to choose a sufficiently large pentagon, to insure that the five points we wish to produce lie within the region common to the five lenses (see Figure 5). A total of five folds will suffice for the pentagon. \square

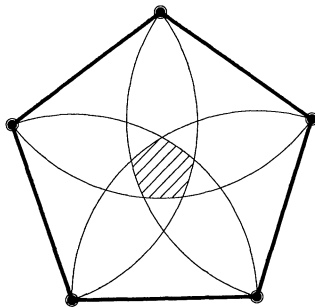


Figure 5. The Pentagon

Theorem 5. *Every six points are the images of the vertices of a regular hexagon under at most seven folds.*

Proof: The procedure is similar to that for the pentagon. Here the lenses determined by each vertex between the two adjacent vertices meet in only one point. Choose a hexagon sufficiently large to contain all six required points within a disc inside a single lens, between the center and some vertex a . Fold the opposite vertex b to the center b' . Label one of the points nearest to the center inside the lens b'' , and label that nearest to a as a' . Now fold all four remaining vertices from the hexagon to the four unlabeled points by folding, first, those vertices formerly adjacent to b , and then those adjacent to a . (Lemma 1 guarantees that this will not disturb a or b' .) Finally, fold a to a' and b' to b'' . This procedure requires seven folds. \square

Lemma 3. *Every regular n -gon can be transformed into a set of n distinct collinear points by a finite sequence of folds. At most $\frac{1}{2}n$ folds are required when n is even, and at most $\frac{1}{2}(n - 1)$ folds are required when n is odd.*

Proof: For n even, choose a suitable line L_1 through the center of the n -gon so that folding about L_1 produces n distinct points. (There are infinitely many lines through the center, and only finitely many of them fail to have this property.) “Stratifying” the images in layers parallel to L_1 , and folding each layer to the one below will collapse all points to a collinear set, as shown in Figure 7a. Observe that for n even, the complete process requires $\frac{1}{2}n$ folds. For n odd, the polygon may be easily stratified in horizontal layers and collapsed to a collinear set by orienting the polygon so that one side is horizontal, as in Figure 7b. For n odd, this process requires $\frac{1}{2}(n - 1)$ folds. \square

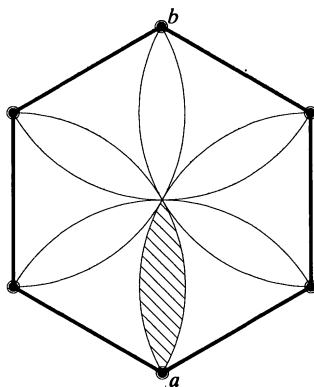


Figure 6. The Hexagon

Lemma 4. *Let $\{a_1, \dots, a_n\}$ be a set of n points contained in an open disc of diameter d . If $\{x_1, \dots, x_n\}$ is a collinear set of n points with $x_1 = a_1$ and $|\overline{x_i x_j}| \geq 2d$ (for $j = i + 1, i = 1, 2, \dots, n - 1$) then there is a sequence of $(n - 1)$ folds which transforms $\{x_1, \dots, x_n\}$ into $\{a_1, \dots, a_n\}$.*

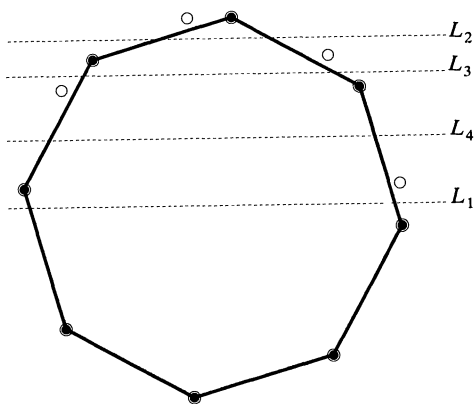


Figure 7a

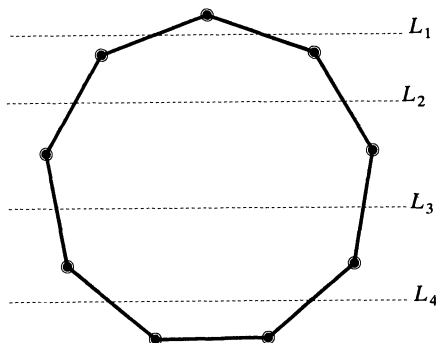


Figure 7b

Proof: We proceed by induction on n . If $n = 2$ the statement is true, since a_2 lies in the circle with center at $x_1 = a_1$, which passes through x_2 . If we assume the assertion of the Lemma to be true for n with $2 \leq n \leq k$, observe that we may fold x_1, \dots, x_k to a_1, \dots, a_k with $(k - 1)$ folds, and the last such fold takes x_{k+1} to some point p at a distance at least $2d$ from x_k . It follows that the circle of diameter d which contains $\{a_1, \dots, a_n\}$ lies inside the circle through p with center at a_k , and the fold along the perpendicular bisector of $\overline{a_{k+1}p}$ takes p to a_{k+1} while leaving a_1, \dots, a_k fixed. Thus k folds suffice to transform x_1, \dots, x_{k+1} into a_1, \dots, a_{k+1} . \square

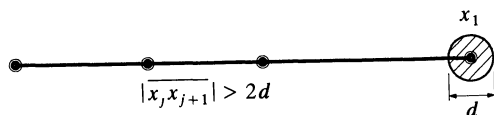


Figure 8. Folding n -collinear points to a small set.

Theorem 6. *Every set of n points is the image of a suitable regular n -gon, under at most finitely many folds. For n even the number of folds does not exceed $\frac{1}{2}(3n - 2)$, and for n odd the number of folds does not exceed $\frac{1}{2}(3n - 3)$.*

Proof: Let S be a set of n points and assume $S \subseteq C$, a circle of diameter d . Choose a regular n -gon P sufficiently large so that the minimum distance between x_1, \dots, x_n (the collinear points constructed from Lemma 3) is at least $2d$. Locate the polygon in the plane so that x_1 coincides with one member of S . Now apply Lemma 4 to transform these collinear points into S . The total number of folds required is then given by Lemmas 3 and 4. For n even, the number of folds does not exceed $\frac{1}{2}n + (n - 1)$ and for n odd, the number of folds does not exceed $\frac{1}{2}(n - 1) + (n - 1)$. \square

Observe that Theorem 6 has a more general analog which may be proved easily using the techniques employed in Lemmas 3 and 4. If R is a set of distinct points,

then every set of n points is the image of a suitable R' similar to R , under at most $2(n - 1)$ folds. At most $(n - 1)$ folds are required to produce a collinear set from R , and a suitably large copy R' can be reduced to the given set by at most $(n - 1)$ additional folds.

5. AN OPEN PROBLEM. For $n = 4, 5, 6, 7, \dots$, it remains an intriguing problem to determine the least number of folds required to produce arbitrary n -point configurations from the vertices of a suitable n -gon. In particular, it would be nice to settle the question whether or not two folds suffice in general to produce four arbitrary points from the square.

ACKNOWLEDGMENT. The authors wish to thank the referee for a suggested improvement to Lemma 3 and for drawing our attention to the observation which follows Theorem 6.

REFERENCES

1. D. Dorninger and W. Timischl, Geometrical Constraints on Bennetts' predictions of Chromosome order, *Heredity* 58 (1987), 321–325.
2. D. Dorninger and G. Hasibeder, Reconstruction of Deformed Regular Polygons: a Problem Arising from Cell Biology, *Demonstratio Mathematica* 21 (1988), 2, 559–563.

Department of Mathematics & Statistics
The University of Calgary
Calgary, Alberta
Canada T2N 1N4
psabinin@acs.ucalgary.ca
mgstone@acs.ucalgary.ca

A strict separation must be maintained between physics and mathematics. Physics must remain quite independent; it must use all its powers of love, respect, and reverence to find its way into nature and the sacred life of nature irrespective of what mathematics does. The latter, on the other hand, must declare itself independent of all externalities, take its own path of intellect, and develop in a purer way than it now does in working with the physical world to gain something from it or impose something on it.

Goethe Scientific Studies, translated by Douglas Miller.

Isometries of the Plane

David A. Singer

The purpose of this note is to develop an elementary proof of the classification of isometries of the Euclidean and hyperbolic planes. The proof is constructive and elementary; it illustrates the distinction between the two geometries by presenting a Euclidean construction which in the non-Euclidean case breaks down at the last step, revealing the horocycle rotation. The proof is developed in the framework of absolute geometry (geometry without the parallel postulate) as far as this can be pushed.

The basic result concerning rigid motions is the following, which can be found in [1, p. 46] for the Euclidean plane:

Theorem 1. *Every rigid motion of the “plane” is of one of the following types:*

- i) *Rotations about a fixed point P ;*
- ii) *Translations in the direction of a line l ;*
- iii) *Reflection across a line l ;*
- iv) *Glide-reflections along a line l .*

The first type has P as a fixed point; the others have l as an invariant line (that is, $T(l) = l$).

Note. The quotation marks above are to remind the reader that in order for this to really be a theorem, one has to be sure of the meaning of the word “plane” in the statement.

Before reading the proof of this result, it is helpful to consider the effect of composing two isometries. The theorem implies that the composition of two rotations must be a rotation or a translation. (It can't be a reflection or glide-reflection.) This is not intuitively obvious, although it is hard to visualize a counter-example.

To prove the theorem, we need a couple of elementary facts about rigid motions. First of all, given three points determining a triangle, there is a unique rigid motion which carries these three points to a specified congruent triangle. (This might well be taken as the definition of congruence.)

Given three non-collinear points A, B, C and two points A', B' with lengths $|AB| = |A'B'|$, there are exactly two rigid motions which carry A to A' and B to B' , determined by the two choices of C' making $\triangle ABC$ congruent to $\triangle A'B'C'$.

Finally, from these observations we see that if an isometry T of the plane has three non-collinear fixed points, it is the identity; if it has two fixed points it is either the identity or the reflection through the line determined by these fixed points.

Now Let T be an isometry of the “plane.” (The quotation marks are to emphasize that I am not going to say whether this is the Euclidean plane or the

hyperbolic plane until absolutely forced to do so. The pedagogical advantage of this approach is that it dramatizes the need for precise and rigorous arguments.) Choose any point x_0 in the plane for which $T(x_0) \neq x_0$. Let $x_1 = T(x_0)$ and let $x_2 = T(x_1)$.

It may happen that $x_2 = x_0$. In this case the line l' determined by x_0 and x_1 is invariant under T and the midpoint A of the segment from x_0 to x_1 is a fixed point. Choosing any point B on the line l through A perpendicular to l' , it is immediate that l is also invariant and that T is either a reflection through l (if B is fixed) or rotation through 180° around A (if $T(B) = B'$ is on the opposite side of l'). [See Figure 1.]

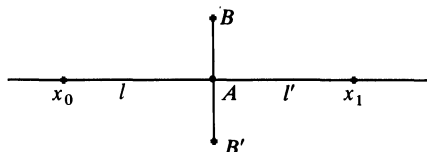


Figure 1

Note that choosing x_0 anywhere in the plane yields $x_2 = x_0$ in both of these cases. Thus we have classified the involutions of the plane ($T^2 = \text{identity}$), and we can see that if $x_2 = x_0$, then the choice is no accident!

Now we may assume that x_0 , x_1 , and x_2 are distinct points. If they are collinear, determining a line l , then T must either be the glide reflection along l or the translation along l carrying the points x_0 and x_1 to the points x_1 and x_2 . (There can only be two rigid motions carrying x_0 to x_1 and x_1 to x_2 , and we have named two.)

We are reduced to the expected situation, namely that x_0 , x_1 , and x_2 form a triangle. Let A be the midpoint of $\overline{x_0x_1}$ and B the midpoint of $\overline{x_1x_2}$. Let $x_3 = T(x_2)$ and let C be the midpoint of $\overline{x_2x_3}$. We have $B = T(A)$ and $C = T(B)$. If A , B , and C are collinear (Figure 2), then inspection of the figure shows that T must be a glide-reflection along this line (a reflection followed by a translation along the reflecting line.) Conversely, if T is a glide reflection then this collinearity always occurs.

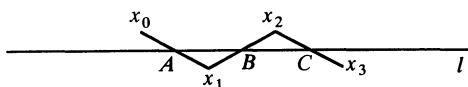


Figure 2

We are down to one final situation, in which the points A , B and C are non-collinear. To show that T is a rotation, we must find the fixed point. But it is clear where to look for it! Construct the perpendicular l to $\overline{x_0x_1}$ through A and the perpendicular l' to $\overline{x_1x_2}$ through B . The map T takes l to l' . If O is the point of intersection of l and l' , then O must be a fixed point and T a rotation around O . (See Figure 3.)

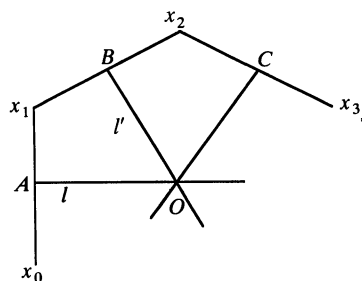


Figure 3

Up until this very last sentence, the reasoning above has been equally valid in the Euclidean case and the hyperbolic case. However, *the existence of O requires the parallel postulate*. Indeed, this illustrates the danger of arguing by pictures. The flaw in the argument lies in the assumption that Figure 3 must look the way it does.

Before formulating the correct theorem for the hyperbolic case, it will be helpful to review the behavior of lines. Recall that in hyperbolic geometry there are *three* ways a pair of lines l and l' can look. They can intersect at some point P . They can be “parallel,” diverging from each other in one direction while approaching each other asymptotically in the other direction. Such lines are said to meet at an “ideal point,” which can be thought of as an “endpoint at infinity” of the line. When the hyperbolic plane is modeled by the Poincaré disc, these ideal points form the boundary circle.

Finally, they can be “hyperparallel,” diverging from each other in both directions. Such lines have the property that they have a common perpendicular, which represents the shortest line segment between the two lines. Given a point A not on a line l , there are infinitely many lines through A not meeting l ; of these, two are parallel and the others are hyperparallel. (See [2], pp. 179–187, [3], p. 156, for information about parallels and common perpendiculars.)

Theorem 2. *Every rigid motion of the hyperbolic plane is one of the following types:*

- i) Rotations about a fixed point P ;
- ii) Translations in the direction of a line l ;
- iii) Reflection across a line l ;
- iv) Glide-reflections along a line l ;
- v) Horocycle rotations.

The first type has P as a fixed point; the second and third have l as an invariant line and have two ideal points fixed (the “endpoints” of l). The fifth kind has one ideal point fixed; it has no invariant lines.

To prove this result by completing the earlier argument, we need to do some work with hyperparallel lines.

Consider the rays of l and l' on the side of the angle $\angle Ax_1B$. Let Z_0 be the point on l at the foot of the perpendicular from B to l . Imagine a moving point Z on l , sliding along (away from A) starting at Z_0 . (See Figure 4.) As Z moves out along l the perpendicular through Z meets l' at a moving point W . If Z_2 is farther out than Z_1 , then angle $\angle Z_2W_2B$ is smaller than $\angle Z_1W_1B$ (since the sum of the angles in the quadrilateral $Z_1Z_2W_1W_2$ is *less* than four right angles).

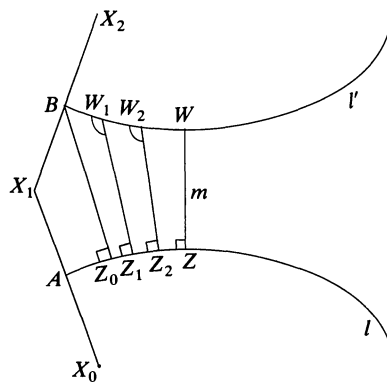


Figure 4

It is possible that for some choice of Z the perpendicular m through Z is also perpendicular to l' at W . In that case, a symmetry argument shows that $|AZ| = |BW|$. (I leave this as an exercise.) Therefore, translation along m will carry l to l' and thus T is that translation. (This reveals the remarkable fact that in the hyperbolic case, a translation has exactly one invariant line; it is characterized as the unique common perpendicular to the lines l and l' .)

If, however, the lines l and l' are parallel, then no such common perpendicular exists. Instead, the two lines approach each other, “meeting” at an ideal point. Since T carries l to l' , this ideal point is “fixed” by the map. More simply, if l'' is a common parallel to l and l' , then $T(l'')$ must also be a common parallel, so that the family of (one-sided) parallels to l are invariant under T . T is therefore a “rotation” about the ideal point. (See [1], p. 269, where “parallel displacement” is discussed.) It is an elementary (but interesting!) exercise to show that no line is fixed under T in this case.

This theorem may be used as a jumping-off point for further exploration. For example, similar arguments will show that in the (double) elliptic case, T is always a rotation. What are the analogous theorems in the 3-dimensional case?

Another direction in which fruitful exploration is available is through the notion of an *orbit*. If T is an isometry of the plane and $x = x_0$ is any point, then the orbit of x is the sequence of points $x_0, x_1, x_2, x_3 \dots$ defined by $x_{n+1} = T(x_n)$. In the Euclidean plane, such an orbit consists either of one point, two distinct points, finitely many points equally spaced on a circle, or infinitely many points lying on a line or circle or on two lines. What can be said about the curves in hyperbolic space (or in higher dimensions) on which the orbits lie?

REFERENCES

1. H. S. M. Coxeter, *Introduction to Geometry*, John Wiley and Sons, 1961.
2. R. Faber, *Foundations of Euclidean and Non-Euclidean Geometry*, Marcel Dekker, Inc., 1983.
3. M. J. Greenberg, *Euclidean and Non-Euclidean Geometries*, W. H. Freeman and Co., 1974.
4. D. Hilbert and S. Cohn-Vossen, *Geometry and the Imagination*, Chelsea, 1952. pp. 59–62 use similar arguments to do part of Theorem 1.

Department of Mathematics
Case Western Reserve University
Cleveland, OH 44106-7058
das5@po.cwru.edu

NOTES

Edited by: John Duncan

One More Construction Which Is Impossible

V. A. Geyler

The purpose of this note is to present a new “elementary” geometrical problem which cannot be solved if one is required to use a straightedge and compass only. At first glance it seems very surprising that a problem like this needs more advanced tools.

Problem. *To construct a secant of a given circle which divides the area of the circle into two commensurate but unequal parts.*

That is, let a circle of unit radius and a positive rational number $q \neq 1$ be given (see Figure 1). We want to construct a secant AB such that the ratio of the areas of the two complementary segments APB and $AP'B$ is equal to q . Recall that a (circular) segment is the region bounded by an arc and the subtending chord.

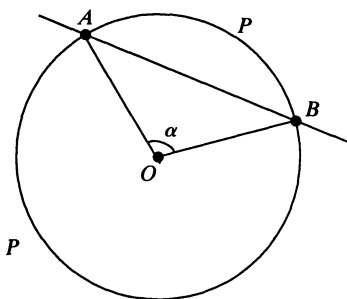


Figure 1

We will show that this construction is impossible if a straightedge and compass are the only tools which we may use. We precede the proof with a couple of definitions and facts which will be needed. A complex number is *algebraic* if it is a zero of a polynomial with integer coefficients. Numbers which are not algebraic are called *transcendental*. The set of all algebraic numbers is a field. This implies, in particular, that for any algebraic number a and any natural number k the number a^k is also algebraic. Recall also that each constructible number, i.e., a number which can be constructed using a straightedge and compass only, is algebraic. A crucial result for us is the following theorem due to Lindemann. If $x \neq 0$ is an algebraic number, then e^x is a transcendental number. We refer to [2, 3] for unexplained terminology and details.

Proof of our claim: We use the notation introduced in Figure 1. Assume, contrary to what we claim, that there exists a rational number n/m , where m, n are positive integers with $1 \leq n < m$, for which we can construct, with straightedge and compass, a secant AB of the circle which divides the area in the ratio $n : m$. Let us denote by A_1 and A_2 the areas of the segments APB and $AP'B$ respectively, and by α the central angle AOB which is assumed to be less than π . The constructibility of the secant AB implies immediately that the numbers $\sin(\alpha/2)$ and $\cos(\alpha/2)$ are constructible, and therefore $\sin \alpha = 2 \sin(\alpha/2) \cos(\alpha/2)$ is algebraic (even constructible).

On the other hand, we have that $2A_1 = \alpha - \sin \alpha$ and $2A_2 = 2\pi - \alpha + \sin \alpha$, and consequently

$$A_1/A_2 = (\alpha - \sin \alpha)/(2\pi - \alpha + \sin \alpha) = n/m.$$

This implies that $(m+n)\alpha - 2\pi n = (m+n)\sin \alpha$. Thus, the number $x = (m+n)\alpha - 2\pi n$ is algebraic, and clearly $x \neq 0$ as $\sin \alpha > 0$. By Lindemann's theorem $e^{ix} = e^{i(m+n)\alpha}$ is a transcendental number, and hence $e^{i\alpha}$ is a transcendental number, too. Since the numbers $\cos \alpha$ and $\sin \alpha$ are either both algebraic or both transcendental, and since $e^{i\alpha} = \cos \alpha + i \sin \alpha$, we can conclude that $\sin \alpha$ is a transcendental number, a contradiction.

Remark 1. Lindemann's theorem is a very deep result, which seems to be much deeper than the problem it is applied to. Therefore it would be interesting to find out if there exists a more elementary proof of the fact proved above without using Lindemann's theorem.

Remark 2. Here we quickly discuss a few easy (and possibly known) results following from Lindemann's theorem and related to the ideas above. For brevity we use the word "construct" to mean "construct with ruler and compass."

First of all, we cannot construct a nonzero angle α whose radian measure is algebraic. (Indeed, a central angle α so constructed on the unit circle determines a chord with length $2 \sin(\alpha/2)$ and the latter number is algebraic. However, as we showed above $[\alpha \text{ algebraic}] \Rightarrow [2 \sin(\alpha/2) \text{ transcendental}]$, a contradiction.)

It follows that we cannot construct a sector (area $= \alpha/2$) or a segment (area $= (\alpha - \sin \alpha)/2 = [\text{transcendental}] - [\text{algebraic}]$) of the unit circle whose area is a nonzero algebraic number. In other words, we can state the well known impossibility of squaring the circle in a more general way, by saying that it is impossible to square a nonzero constructible sector or segment of the unit circle.

Finally, following a comment by the referee, we would like to mention a very interesting theorem of Newton which seems quite relevant. *Each smooth oval, in particular a circle, is algebraically non-integrable. This means that there is no non-zero polynomial P satisfying $P(S, a, b, c) = 0$, where S is the area of the segment cut off by the straight line $ax + by = c$.* We refer to [1, pp. 84–85] for details regarding this fact.

In conclusion the author would like to thank the referee for his suggestions, and also M. Frantz and J. Sarkar for their help in preparing the manuscript.

REFERENCES

1. V. I. Arnol'd, *Huygens and Barrow, Newton and Hooke: pioneers in mathematical analysis and catastrophe theory from evolvents to quasicrystals*, Birkhäuser Verlag, Basel · Boston · Berlin, 1990.

2. J. B. Fraleigh, *A First Course in Abstract Algebra*, 4th ed., Addison-Wesley, 1989.
3. A. G. Kurosh, *A Course in Higher Algebra*, 10th ed., Nauka, Moscow, 1971 (in Russian).

Department of Mathematics
Mordovian State University
430000 Saransk
Russia

An Inductive Proof of a Mixed Arithmetic-Geometric Mean Inequality

Takashi Matsuda

Let x_1, \dots, x_n be positive real numbers. The arithmetic-geometric mean inequality is

$$y_n = \frac{x_1 + \dots + x_n}{n} \geq \sqrt[n]{x_1 \dots x_n} = z_n \quad (1)$$

and a mixed arithmetic-geometric mean inequality is

$$\alpha_n = \frac{z_1 + \dots + z_n}{n} \leq \sqrt[n]{y_1 \dots y_n} = \gamma_n. \quad (2)$$

Equality holds if and only if $x_1 = \dots = x_n$.

A proof of (2) was given by Kiran Kedlaya [1]; it was combinatorial. In this note we propose to give an inductive proof of (2) that uses a little analysis. The present proof is suitable for the advanced undergraduate student.

We shall first give a proof of the following lemma.

Lemma. *Let*

$$S := \{(u_1, \dots, u_{m+1}) : u_1 + \dots + u_{m+1} = (m+1)a, u_1 > 0, \dots, u_{m+1} > 0\},$$

where a is a positive constant and let

$$f(u_1, \dots, u_{m+1}) := \frac{1}{m+1} \left(u_1 + \frac{u_2^2}{u_1} + \dots + \frac{u_{m+1}^{m+1}}{u_m^m} \right) \left(\frac{u_1 + \dots + u_m}{m} \right)^m, \\ (m = 1, 2, \dots).$$

Then

$$f(u_1, \dots, u_{m+1}) \geq a^{m+1} \quad \text{for every } (u_1, \dots, u_{m+1}) \in S. \quad (3)$$

Equality holds if and only if $u_1 = \dots = u_{m+1}$.

Proof: If $m = 1$, then we have

$$f(u_1, u_2) = \frac{u_1^2 + (2a - u_1)^2}{2} \geq a^2.$$

If $m \geq 2$, then let

$$g(u_1, \dots, u_{m+1}) := \frac{1}{m+1} \left(2u_2 + \frac{u_3^3}{u_2^2} + \dots + \frac{u_{m+1}^{m+1}}{u_m^m} \right) \left(\frac{u_1 + \dots + u_m}{m} \right)^m.$$

Since $u_1 + \frac{u_2^2}{u_1} \geq 2u_2$, we have

$$f(u_1, \dots, u_{m+1}) \geq g(u_1, \dots, u_{m+1}). \quad (4)$$

Let p be a positive integer and let

$$\begin{aligned} S_p &:= \left\{ (u_1, \dots, u_{m+1}) : u_1 + \dots + u_{m+1} \right. \\ &\quad \left. = (m+1)a, u_1 \geq \frac{a}{p+1}, \dots, u_{m+1} \geq \frac{a}{p+1} \right\}. \end{aligned}$$

Since g is continuous and S_p is a compact set, g has a maximum and a minimum value on S_p . For $u_1 > 0, u_2 > 0, \dots, u_{m+1} > 0$, let

$$h(u_1, \dots, u_{m+1}) := g(u_1, \dots, u_{m+1}) - \lambda[u_1 + \dots + u_{m+1} - (m+1)a].$$

We proceed by computing

$$\frac{\partial h}{\partial u_1}, \dots, \frac{\partial h}{\partial u_{m+1}} \quad \text{and} \quad \frac{\partial h}{\partial \lambda}$$

and setting each of these expressions equal to zero. We obtain

$$\frac{\partial h}{\partial u_1} = \frac{1}{m+1} \left(2u_2 + \dots + \frac{u_{m+1}^m}{u_m^m} \right) \left(\frac{u_1 + \dots + u_m}{m} \right)^{m-1} - \lambda = 0, \quad (5)$$

$$\frac{\partial h}{\partial u_2} = \frac{2}{m+1} \left(1 - \frac{u_3^3}{u_2^3} \right) \left(\frac{u_1 + \dots + u_m}{m} \right)^m + \frac{\partial h}{\partial u_1} = 0, \quad (6)$$

$$\begin{aligned} \frac{\partial h}{\partial u_k} &= \frac{k}{m+1} \left(\frac{u_k^{k-1}}{u_{k-1}^{k-1}} - \frac{u_{k+1}^{k+1}}{u_k^{k+1}} \right) \left(\frac{u_1 + \dots + u_m}{m} \right)^m + \frac{\partial h}{\partial u_1} = 0, \\ &\quad (k = 3, \dots, m), \end{aligned} \quad (7)$$

$$\frac{\partial h}{\partial u_{m+1}} = \frac{u_{m+1}^m}{u_m^m} \left(\frac{u_1 + \dots + u_m}{m} \right)^m - \lambda = 0, \quad (8)$$

$$\frac{\partial h}{\partial \lambda} = -(u_1 + \dots + u_{m+1}) + (m+1)a = 0. \quad (9)$$

From (5), (6) and (7), we have

$$u_2 = \dots = u_m = u_{m+1}.$$

If we set $u_2 = \dots = u_{m+1} = t$, then from (5) and (8), we have $u_1 = t$ and using (9), we get

$$u_1 = u_2 = \dots = u_{m+1} = a.$$

Thus, the system of equations

$$\frac{\partial h}{\partial u_1} = \dots = \frac{\partial h}{\partial u_{m+1}} = \frac{\partial h}{\partial \lambda} = 0$$

has one and only one solution (a, \dots, a) and $(a, \dots, a) \in S_p$. This critical point obtained by the Lagrange multiplier rule must give the maximum or the minimum.

Thus,

$$g(a, \dots, a) = a^{m+1}$$

is the maximum value or the minimum value.

Let $u_1 = \dots = u_m = x$, x some positive number. Then $u_{m+1} = (m+1)a - mx$ and

$$g(x, \dots, x, (m+1)a - mx) = \frac{mx^{m+1} + [(m+1)a - mx]^{m+1}}{m+1}.$$

If $a/(p+1) < x < a + p/m(p+1)a$, then $(x, \dots, x, (m+1)a - mx) \in S_p$. Also, if $x \neq a$, then

$$(x, \dots, x, (m+1)a - mx) \neq (a, \dots, a, a).$$

An easy application of differential calculus establishes the following inequality

$$g(x, \dots, x, (m+1)a - mx) > a^{m+1}, \left(\frac{a}{p+1} < x < a + \frac{p}{m(p+1)}a, x \neq a \right). \quad (10)$$

Hence, a^{m+1} is the minimum value of g on S_p . Since $S_p \uparrow S$ as p becomes arbitrarily large, we have

$$g(u_1, \dots, u_{m+1}) \geq a^{m+1} \quad \text{for every } (u_1, \dots, u_{m+1}) \in S.$$

From (4), we therefore also have

$$f(u_1, \dots, u_{m+1}) \geq a^{m+1} \quad \text{for every } (u_1, \dots, u_{m+1}) \in S.$$

It is clear that equality holds if and only if $u_1 = \dots = u_{m+1}$.

This proves the lemma. ■

Proof of (2): Inequality (2) is equivalent to

$$\alpha_n^n = \left(\frac{z_1 + \dots + z_n}{n} \right)^n \leq y_1 \dots y_n = \gamma_n^n. \quad (11)$$

Let $a = \alpha_{m+1}$ in the lemma; then $(z_1, \dots, z_{m+1}) \in S$. Furthermore, it is quite easy to see that

$$f(z_1, \dots, z_{m+1}) = y_{m+1} \left(\frac{z_1 + \dots + z_m}{m} \right)^m.$$

Let $m = 1$. Using the lemma, we have

$$\alpha_2^2 \leq y_2 z_1 = y_1 y_2 = \gamma_2^2.$$

Evidently, equality holds if and only if $x_1 = x_2$. Hence (11) is true for $n = 2$.

Mathematical induction now readily gives the general result (11).

Thus, assume that (11) is valid for n . Let $m = n$. Using the lemma and the induction hypothesis, we have

$$\alpha_{n+1}^{n+1} \leq y_{n+1} \left(\frac{z_1 + \dots + z_n}{n} \right)^n \leq y_1 \dots y_n y_{n+1} = \gamma_{n+1}^{n+1}. \quad (12)$$

Evidently, $x_1 = \dots = x_n$ is equivalent to $z_1 = \dots = z_n$. Hence, equality in the second inequality of (12) holds if and only if $z_1 = \dots = z_n$.

Therefore, equality holds simultaneously in both inequalities of (12) if and only if $z_1 = \cdots = z_{n+1}$, i.e., $x_1 = \cdots = x_{n+1}$.

Thus we have established the mixed arithmetic-geometric mean inequality (2).

REFERENCE

1. Kiran Kedlaya, Proof of a Mixed Arithmetic-Mean, Geometric-Mean Inequality, *Amer. Math. Monthly*, 101 (1994), 355–357.

Department of Mathematics
Kochi Medical School
Okohcho, Nankoku
Kochi 783
Japan

The Ranks of Tournament Matrices

T. S. Michael

1. TOURNAMENT MATRICES. A round-robin tournament is held among n players. Altogether there are $n(n-1)/2$ matches, and we assume that no match ends in a draw. The results are recorded in an n by n tournament matrix $A = [a_{ij}]$ as follows. Label the players with the indices $1, \dots, n$ and define

$$a_{ij} = \begin{cases} 1 & \text{if player } i \text{ defeats player } j, \\ 0 & \text{otherwise.} \end{cases}$$

Thus the tournament matrix A satisfies

$$A + A^T + I = J, \tag{1}$$

where I is the n by n identity matrix and J is the n by n matrix of all 1's. Conversely, any matrix A of 0's and 1's that satisfies (1) is a tournament matrix. The score s_i is the number of players defeated by player i . A tournament is *regular* provided each player has the same score. For $i \neq j$ the *joint score* s_{ij} is the number of players defeated by both player i and player j .

Because each element of A is 0 or 1, we may view A as a matrix over any field. In a note in this *Monthly*, de Caen [1] uses elementary linear algebra to establish a general inequality for the rank of a tournament matrix.

Proposition (de Caen). *If A is an n by n tournament matrix, then $\text{rank}(A) \geq (n-1)/2$ over any field. If A is a regular n by n tournament matrix and $\text{rank}(A) = (n-1)/2$, then the characteristic of the field divides $(n-1)/2$.*

We shall prove de Caen's inequality in a manner that allows us to characterize the case of equality both algebraically and combinatorially. In particular, we shall see that the hypothesis of regularity may be dropped in the second assertion of the Proposition. Here is our main result.

Theorem 1. *Let A be an n by n tournament matrix. Then*

$$\text{rank}(A) \geq (n-1)/2 \quad (2)$$

over any field F . Equality holds if and only if n is odd and $AA^T = O$. Moreover, equality implies that the characteristic of F divides $(n-1)/2$.

In §3 we construct tournament matrices that achieve the lower bound in (2). We remark that ranks of tournament matrices over finite fields may be used in an important combinatorial problem—the testing of block designs for isomorphism [2].

Observe that if A is a tournament matrix, then the (i, i) element of AA^T is equal to the score s_i , while the (i, j) element ($i \neq j$) is equal to the joint score s_{ij} . This observation is a consequence of the row-by-column definition of matrix multiplication and the manner in which tournament matrices are defined. Thus from Theorem 1 we immediately obtain the following combinatorial description of the case of equality in (2).

Corollary. *An n by n tournament matrix A satisfies $\text{rank}(A) = (n-1)/2$ over the field F if and only if n is odd and the characteristic of F divides every score and every joint score of the tournament corresponding to A .*

2. PROOF OF THEOREM 1. By the rank-plus-nullity theorem inequality (2) is equivalent to

$$\text{nullity}(A) \leq (n+1)/2. \quad (2')$$

Equality holds in (2) if and only if equality holds in (2'). We shall work with (2') throughout our argument.

The case $n = 1$ is trivial. Henceforth suppose that $n \geq 2$. Let $\mathbf{u} = (1, \dots, 1)^T$ denote the column vector of n 1's, and let \mathbf{e}_i denote the i th unit coordinate vector ($i = 1, \dots, n$). Then $J\mathbf{u} = n\mathbf{u}$ and $J\mathbf{e}_i = \mathbf{u}$ for $i = 1, \dots, n$. Thus over any field F the only possible nonzero eigenvalue of J is n , and the corresponding eigenspace would be $\langle \mathbf{u} \rangle$.

Let N_A denote the nullspace of A . Suppose that $\mathbf{v} \in N_A \cap N_{A^T}$ and $\mathbf{v} \neq \mathbf{0}$. Then $J\mathbf{v} = (A + A^T + I)\mathbf{v} = \mathbf{v}$ by (1). Hence \mathbf{v} is an eigenvector of J corresponding to the eigenvalue 1. By the preceding comments, $n = 1$ in F , and $\mathbf{v} \in \langle \mathbf{u} \rangle$. Thus the dimension of the subspace $N_A \cap N_{A^T}$ is at most 1. Hence

$$\text{nullity}(A) + \text{nullity}(A^T) \leq n + \dim(N_A \cap N_{A^T}) \leq n + 1.$$

Now $\text{nullity}(A) = \text{nullity}(A^T)$, and therefore inequality (2') is true.

Suppose that equality holds in (2'). Of course, n must be odd. Also, the nullspaces N_A and N_{A^T} have intersection $\langle \mathbf{u} \rangle$ and dimension $(n+1)/2$. Thus there are bases $\{\mathbf{u}, \mathbf{v}_1, \dots, \mathbf{v}_{(n-1)/2}\}$ and $\{\mathbf{u}, \mathbf{w}_1, \dots, \mathbf{w}_{(n-1)/2}\}$ for N_A and N_{A^T} , respectively, whose union spans $F^{(n)}$. To prove that $AA^T = O$ it suffices to prove that AA^T annihilates these two bases. Clearly, $AA^T\mathbf{u} = \mathbf{0}$ and $AA^T\mathbf{w}_i = \mathbf{0}$ for each i . Also, $AJ = O$ and hence by (1)

$$AA^T\mathbf{v}_i = A(J - A - I)\mathbf{v}_i = (AJ - A(A + I))\mathbf{v}_i = -(A + I)A\mathbf{v}_i = \mathbf{0}.$$

for each i . Therefore $AA^T = O$. Let the characteristic of F be p . The sum of the diagonal elements of AA^T is the sum of the scores of the tournament, i.e., $n(n-1)/2$. Hence $AA^T = O$ implies that p divides $n(n-1)/2$, and thus p divides $(n-1)/2$, as $n = 1$ in F .

Conversely, suppose that n is odd and $AA^T = O$. Then the nullspace of A contains the column space of A^T . Thus

$$\text{nullity}(A) \geq \text{rank}(A^T) = \text{rank}(A) = n - \text{nullity}(A),$$

and so $\text{nullity}(A) \geq n/2$. Also, $\text{nullity}(A) \leq (n+1)/2$ by (2'). Because n is odd, the only possibility is that $\text{nullity}(A) = (n+1)/2$. ■

3. DOUBLY REGULAR TOURNAMENTS. A tournament among n players ($n \geq 3$) is *doubly regular* provided the joint scores are all equal, say, to $m-1$, where m is a positive integer. In a doubly regular tournament the set of players defeated by player i yields a regular tournament in which each player defeats $m-1$ others ($i = 1, \dots, n$). Thus a doubly regular tournament is regular with each score equal to $2m-1$ and with $4m-1$ players. Let A_m be a doubly regular $4m-1$ by $4m-1$ tournament matrix. Then we have shown that

$$A_m A_m^T = mI + (m-1)J. \quad (3)$$

Theorem 2. Let A_m be a doubly regular $4m-1$ by $4m-1$ tournament matrix, and let F be a field of characteristic p , where p divides m . Write $n = 4m+1$. Then the n by n tournament matrix

$$A = \left[\begin{array}{ccc|cc} & & & 1 & 0 \\ & & & \vdots & \vdots \\ & A_m & & \vdots & \vdots \\ & & & 1 & 0 \\ \hline 0 & \cdots & 0 & 0 & 0 \\ 1 & \cdots & 1 & 1 & 0 \end{array} \right]$$

has rank $(n-1)/2$ over F .

Proof: Over the field of rational numbers

$$AA^T = \left[\begin{array}{ccc|cc} & & & 0 & 2m \\ & & & \vdots & \vdots \\ & m(I+J) & & \vdots & \vdots \\ & & & 0 & 2m \\ \hline 0 & \cdots & 0 & 0 & 0 \\ 2m & \cdots & 2m & 0 & 4m \end{array} \right]$$

by (3). Hence $AA^T = O$ over F . Also, n is odd. Therefore $\text{rank}(A) = (n-1)/2$ by Theorem 1. ■

Doubly regular $4m-1$ by $4m-1$ tournament matrices exist for infinitely many values of m . (See §7 of [3] for a plethora of constructions.) Thus Theorem 2 shows that equality holds infinitely often in Theorem 1.

REFERENCES

1. D. de Caen, The ranks of tournament matrices, *Amer. Math. Monthly*, 98 (1991), 829–831.
2. T. S. Michael, The p -ranks of skew Hadamard designs, to appear in *J. Comb. Theory, Ser. A*.
3. J. Seberry and M. Yamada, Hadamard matrices, sequences, and block designs, in *Contemporary Design Theory* (J. H. Dinitz and D. R. Stinson, eds.), Wiley, New York, 1992.

Mathematics Department
United States Naval Academy
Annapolis, MD 21402
tsm@sma.usna.navy.mil

On Some Applications of Fibonacci Numbers

David L. Ranum

The Fibonacci numbers (0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, ...) as defined by the familiar recurrence

$$\begin{aligned}F_0 &= 0, \\F_1 &= 1, \\F_i &= F_{i-1} + F_{i-2} \text{ for all } i \geq 2\end{aligned}\tag{1}$$

are an important part of mathematics legend. Although they are no longer utilized as a population model for rabbits, the sequence often emerges in very unique and interesting places. In mathematics textbooks, their relationship (Eq. 2) to the golden ratio and its conjugate (shown as an approximation in Eq. 3) appears as an initial nontrivial example of proof by induction.

$$F_i = (\Phi^i - \Psi^i) / \sqrt{5} \text{ where } \Phi = 1.61803 \dots \text{ and } \Psi = -0.61803 \dots \tag{2}$$

$$F_i \approx \Phi^i / \sqrt{5} \tag{3}$$

The computer science community also considers these values to have great importance. It is almost a guarantee that in the first introduction to recursion, a student will be subject to this very same recurrence as a recursive function to compute the n th Fibonacci number. However, this interest goes much further than a simple algorithmic example. In fact, the Fibonacci numbers and their associated properties have been used often in the development and analysis of data structures and algorithms.

BINARY TREES AND BINARY SEARCH TREES. A **binary tree** is a data structure defined on a finite set of nodes (values) that is either:

- (1) empty, or
- (2) consists of a root node with the remaining nodes divided into two disjoint sets (called the left and right subtrees) each of which is a binary tree.

Figure 1 shows an example of a binary tree constructed from seven nodes. The **root** of the entire tree is the node with value 'E.' The node containing 'B' is known as the **left child** of 'E' and similarly the node containing 'G' is known as the **right child** of 'E.' Note that the left child of 'E' also serves as the root of the left subtree from the recursive definition (likewise for the right). The node labeled 'C' is an example of a **leaf** node as it serves as a root for 2 empty subtrees. The **height** of the tree is calculated as the number of levels (or generations) within the tree. In Figure 1 then, the height is said to be 4.

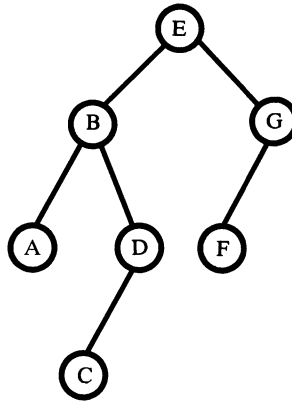


Figure 1. A binary tree.

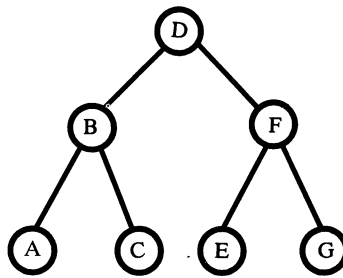


Figure 2. A perfectly balanced, full tree.

Although the binary tree is restricted somewhat, it is apparent that a large number of tree structures are possible given a particular set of node values. At one extreme, Figure 2 shows a **perfectly balanced** tree where each node except for the leaves has both a left and a right child and all leaves appear at the same level. A tree such as this is said to be **full** because it contains the maximum number of nodes for a given height. At the other extreme, Figure 3 shows a worst case **degenerate** tree where each node has only 1 child except for the single leaf.

Claim 1. For a binary tree T of height H : (a) the minimum number of nodes is H and (b) the maximum number of nodes is $2^H - 1$.

Proof: (a) By definition, the tree must have at least 1 node on each level. Therefore, in the minimum case, the tree could contain as few as H nodes and still satisfy this property.

(b) In the maximum case the proof is by induction on the height of the tree.

(base) $H = 1$: If T is of height 1 then there is only one node which must be the root node.

(hyp) $H = i$: number of nodes is $2^i - 1$.

(induction) $H = i + 1$: By definition, a tree of height $i + 1$ must consist of a root node and two subtrees each of height i . By the inductive hypothesis, the

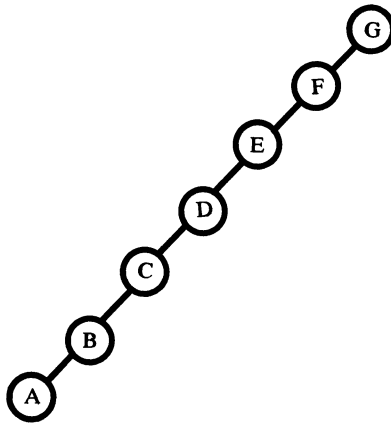


Figure 3. A degenerate binary tree.

subtrees must each have $2^i - 1$ nodes so that total number is $1 + (2^i - 1) + (2^i - 1) = 2(2^i) - 1 = 2^{i+1} - 1$. ■

One common application for any data structure is that of **search**. The problem is to look into the structure and find a selected node value. For this purpose, a special type of binary tree, called a **binary search tree** is often enlisted. A binary search tree is defined as a binary tree where the structural restrictions are related to the node values. In particular, the binary search tree is constructed such that the left child of any node must be less than that node, and the right child of any node must be greater. Which respect to the recursive definition, all the values in the left subtree must be less than the root, and all the values in the right subtree must be greater. This relationship is then required to hold for all subtrees.

If we consider the node values in Figure 1 as having standard alphabetical ordering, then this binary tree is in fact a binary search tree. Likewise, Figures 2 and 3 also satisfy the binary search tree property on the same values, albeit with different shapes. In contrast, Figure 4 is not a binary search tree as the value *B* appears to the left of *A* (to name just one problematic pair).

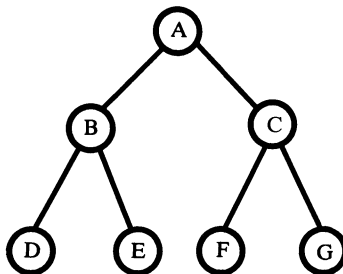


Figure 4. A binary tree lacking the binary search tree properties.

The search algorithm (given below in Pascal-like pseudocode) for binary search trees can be written as a simple recursive procedure which is derived directly from the recursive definition of binary trees.

```

procedure BstSearch (Tree, Value);
{Given a binary search Tree and a Value, search the structure for an occurrence of
the value}
begin
if Tree is empty then {not found}
otherwise
    if Tree.RootValue = Value then {found it}
    otherwise
        if Value < Tree.RootValue then
            BstSearch (Tree.LeftSubtree, Value){must be in the
                                                    left subtree}
        otherwise BstSearch(Tree.RightSubtree, Value){must be in the
                                                        right subtree}
end.

```

In the worst scenario, as in the degenerate tree case of Figure 3, this algorithm will compare against every value in the tree, in essence performing an exhaustive, sequential search upon the elements. However, in a best case situation such as the perfectly balanced case of Figure 2, the algorithm will compare against a maximum of only 3 values as it divides the search space in half every time it performs a compare. These ideas are summarized in the following claim:

Claim 2. Let T be a binary search tree containing n node values. (a) The worst case performance (in terms of compares necessary) for BstSearch is proportional to n . (b) The best case performance for BstSearch is proportional to $\log_2(n)$.

Proof: It can be seen from the above procedure that each compare occurs on a new level within the tree. Therefore, the number of compares is equal to the number of levels or height of the tree. Using claim 1, in the degenerate case there is only one value on each level and therefore with n values there will be n compares. In the perfectly balanced case, a tree of height H has $2^H - 1$ nodes and therefore if a tree has n nodes, the height of such a tree will be bounded by $\log_2(n + 1)$ making the number of compares proportional to $\log_2(n)$. ■

This elementary analysis of binary search tree performance leads to the conclusion that search is better in the case where the height of the tree is minimized. Unfortunately, it is not possible in the general case to have a tree that is perfectly balanced since only certain numbers of nodes can lead to such a tree. Further, although minimum height trees are possible for any given number of nodes, the construction is time consuming and is usually not cost effective. It is therefore necessary to come to some compromise as to the shape of a tree in terms of performance.

AVL BINARY SEARCH TREES. Given any binary search tree, it may be possible to rearrange the nodes so that the resulting new tree is shorter in height and will therefore offer better performance. However, the work involved in this balancing process can offset the expected gain in search performance. An **AVL binary search tree** (named after Russian mathematicians G. M. Adel'son-Vel'skii and E. M. Landis) is one in which the height is not necessarily minimal but is kept close enough that good performance can still be expected. In particular, an AVL binary search tree is defined as a binary search tree where the heights of all left and right

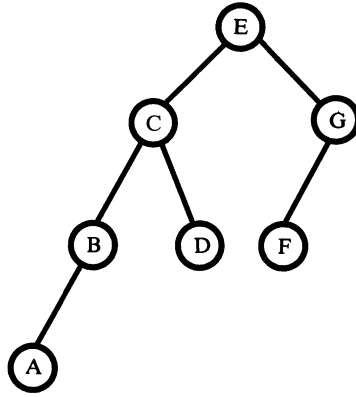


Figure 5. A leftmost AVL binary search tree.

subtrees differ by at most 1. Clearly, the full tree from Figure 2 can be considered AVL since the height difference for any left and right subtree is in fact zero. Figure 1 is also an AVL binary search tree.

What is the expected performance for an AVL binary search tree? It can be seen that the best case for an AVL tree is still that of the perfectly balanced, full tree. Therefore, the number of compares is still logarithmic in the number of nodes in the tree. On the other hand, the worst case will arise when each subtree is as much “out of balance” as possible. Figure 5 shows one such case. Here the AVL tree is worst case in that the height of each left subtree is one greater than the height of its corresponding right subtree. This **leftmost AVL** tree represents one example of a tree which maintains the AVL properties and has the maximum height for this number of nodes. The analysis is not quite so obvious.

Claim 3. The height of a leftmost AVL binary search tree is proportional to the log of the number of nodes in the tree.

Proof: Let N_H be the number of nodes in a leftmost AVL binary search tree of height H . Since the height of the left subtree is one more than the height of the right subtree and both must also be AVL, the following relationship holds:

$$N_H = 1 + N_{H-1} + N_{H-2} \quad \text{where } N_1 = 1 \text{ and } N_2 = 2.$$

This recurrence appears to be very similar to the Fibonacci recurrence stated earlier. In fact, careful observation shows that

$$N_H = F_{H+2} - 1, \quad H \geq 1.$$

By replacing the Fibonacci reference by its golden ratio approximation (Eq. 3), we get

$$N_H \approx \Phi^{(H+2)} / \text{sqrt}(5) - 1.$$

Rearranging terms, taking logs (base 2) of both sides, and solving for H yields

$$\lg(N_H + 1) \approx (H + 2)\lg(\Phi) - (1/2)\lg(5)$$

or

$$H \approx [\lg(N_H + 1) - 2\lg(\Phi) + (1/2)\lg(5)] / \lg(\phi)$$

or

$$H \approx 1.44 \lg(N_H).$$

This result shows that even in worst case, AVL trees still maintain the logarithmic search behavior with respect to the number of nodes in the tree since as before search is related to the height of the tree.

ADDITIONAL APPLICATIONS. The above analysis is interesting not only from a data structures and algorithms point of view but also due to the inclusion of the Fibonacci result. This observation made a seemingly difficult analysis quite modest since once the AVL recurrence was stated in terms of Fibonacci, most of the work was done. The usefulness of the Fibonacci numbers appears elsewhere in the computer science literature as well.

Consider the alphabet consisting of the symbols a and b . A word can then be defined as a sequence of elements from the underlying alphabet. Given this background, it is possible to define the notion of a **Fibonacci Word** as

$$FW_1 = a,$$

$$FW_2 = b,$$

$$FW_{n+2} = FW_{n+1}FW_n \quad \text{for all } n > 2$$

where FW_{n+2} is formed by concatenating FW_n to the right of FW_{n+1} . As examples, $FW_3 = ba$, $FW_4 = bab$, and $FW_5 = babba$. Although these words may seem peculiar, it turns out that they pose a worst case situation for a number of fast pattern matching algorithms [1] which attempt to solve the typical text editing problem of trying to locate a particular sequence of characters in a text file. Interested readers can find additional variations on the Fibonacci theme including **Fibonacci Heaps**, **Fibonacci Merge**, and **Fibonacci Search** in many data structures and algorithms textbooks [2].

REFERENCES

1. D. Knuth, J. Morris, and R. Pratt, Fast Pattern Matching in Strings, *SIAM J. Computing* 6(2), June 1977, pp. 323–350.
2. E. Horowitz, and S. Sahni, *Fundamentals of Data Structures in Pascal*, W. H. Freeman and Company, New York, NY, 1992.

*Department of Computer Science
Luther College
Decorah, IA 52101
ranum@martin.luther.edu*

Abstractness, sometimes hurled as a reproach at mathematics, is its chief glory and its surest title to practical usefulness. It is also the source of such beauty as may spring from mathematics.

—E. T. Bell

THE EVOLUTION OF . . .

Edited by Abe Shenitzer

Mathematics, York University, North York, Ontario M3J 1P3, Canada

Part II. Topology and Abstract Algebra as Two Roads of Mathematical Comprehension*

Unterrichtsblätter für Mathematik und Naturwissenschaften 38, 177-188 (1932). (A lecture in the summer course of the Swiss Society of Gymnasium Teachers, given in Bern, in October 1931.)

Hermann Weyl

Note: The first part of this article appeared in 1995, in the May issue of the *Monthly* (pp. 453). What follows is a short summary of the first part and the concluding part of the article.

SUMMARY OF PART I. Weyl begins by saying that

We are not very pleased when we are forced to accept a mathematical truth by virtue of a complicated chain of formal conclusions and computations, which we traverse blindly, link by link, feeling our way by touch. We want first an overview of the aim and of the road; we want to understand the *idea* of the proof, the deeper context.

and goes beyond this familiar notion of understanding to what he calls modes of understanding. These are ways of looking at mathematics as well as the branches of mathematics associated with them. Two such modes of understanding “have proved, in our time, to be especially penetrating and fruitful. The two are topology and abstract algebra.”

Now the discussion begins to involve some technical matters. Weyl explains what is meant by purely topological investigations of continua, discusses the motives that have led to the development of abstract algebra, and uses “a simple example to show how the same issue can be looked at from a topological and from an abstract-algebraic viewpoint. The (not so simple) example which he considers from the two viewpoints is the theory of algebraic functions of a single variable. ■

After all these general remarks I want to use two simple examples that illustrate the different kinds of concept building in algebra and in topology. The classical example of the fruitfulness of the topological method is Riemann’s theory of

*The original German version of this article is found in vol. 3, pp. 348–358, of the four-volume edition of Hermann Weyl’s collected works published by Springer-Verlag in 1968. The translation is by Abe Shenitzer.

algebraic functions and their integrals. Viewed as a topological surface, a Riemann surface has just one characteristic, namely its connectivity number or genus p . For the sphere $p = 0$ and for the torus $p = 1$. How sensible it is to place topology ahead of function theory follows from the decisive role of the topological number p in function theory on a Riemann surface. I quote a few dazzling theorems: The number of linearly independent everywhere regular differentials on the surface is p . The total order (that is, the difference between the number of zeros and the number of poles) of a differential on the surface is $2p - 2$. If we prescribe more than p arbitrary points on the surface, then there exists just one single-valued function on it that may have simple poles at these points but is otherwise regular; if the number of prescribed poles is exactly p , then, if the points are in general position, this is no longer true. The precise answer to this question is given by the Riemann-Roch theorem in which the Riemann surface enters only through the number p . If we consider all functions on the surface that are everywhere regular except for a single place ρ at which they have a pole, then its possible orders are all numbers $1, 2, 3, \dots$ except for certain powers of p (the Weierstrass gap theorem). It is easy to give many more such examples. The genus p permeates the whole theory of functions on a Riemann surface. We encounter it at every step, and its role is direct, without complicated computations, understandable from its topological meaning (provided that we include, once and for all, the Thomson-Dirichlet principle as a fundamental function-theoretic principle).

The Cauchy integral theorem gives topology the first opportunity to enter function theory. The integral of an analytic function over a closed path is 0 only if the domain that contains the path and is also the domain of definition of the analytic function is simply connected. Let me use this example to show how one "topologizes" a function-theoretic state of affairs. If $f(z)$ is analytic, then the integral $\int_{\gamma} f(z) dz$ associates with every curve a number $F(\gamma)$ such that

$$(\dagger) \quad F(\gamma_1 + \gamma_2) = F(\gamma_1) + F(\gamma_2).$$

$\gamma_1 + \gamma_2$ stands for the curve such that the beginning of γ_2 coincides with the end of γ_1 . The functional equation (\dagger) marks the integral $F(\gamma)$ as an additive path function. Also, each point has a neighborhood such that $F(\gamma) = 0$ for each closed path γ in that neighborhood. I will call a path function with these properties a topological integral, or briefly, an integral. In fact, all this concept assumes is that there is given a continuous manifold on which one can draw curves; it is the topological essence of the analytic notion of an integral. Integrals can be added and multiplied by numbers. The topological part of the Cauchy integral theorem states that on a simply connected manifold every integral is homologous to 0 (not only in the small but in the large), that is, $F(\gamma) = 0$ for every closed curve γ on the manifold. In this we can spot the definition of "simply connected." The function-theoretic part states that the integral of an analytic function is a topological integral in our sense of the term. The definition of the order of connectivity [that we are about to state] fits in here quite readily. Integrals F_1, F_2, \dots, F_n on a closed surface are said to be linearly independent if they are not connected by a homology relation

$$c_1 F_1 + c_2 F_2 + \dots + c_n F_n \sim 0$$

with constant coefficients c_i other than the trivial one, when all the c_i vanish. The order of connectivity of a surface is the maximal number of linearly independent integrals. For a closed two-sided surface the order of connectivity h is always an

even number $2p$, where p is the genus. From a homology between integrals we can go over to a homology between closed paths. The path homology

$$n_1\gamma_1 + n_2\gamma_2 + \cdots + n_r\gamma_r \sim 0$$

states that for every integral F we have the equality

$$n_1F(\gamma_1) + n_2F(\gamma_2) + \cdots + n_rF(\gamma_r) = 0.$$

If we go back to the topological skeleton that decomposes the surface into elementary pieces and replace the continuous point-chains of paths by the discrete chains constructed out of elementary pieces, then we obtain an expression for the order of connectivity h in terms of the numbers s , k and e of pieces, edges and vertices. The expression in question is the well-known Euler polyhedral formula $h = k - (e + s) + 2$. Conversely, if we start with the topological skeleton, then our reasoning yields the result that this combination h of the number of pieces, edges and vertices is a topological invariant, namely it has the same value for “equivalent” skeletons which represent the same manifold in different subdivisions.

When it comes to application to function theory, it is possible, using the Thompson-Dirichlet principle, to “realize” the topological integrals as actual integrals of everywhere regular-analytic differentials on a Riemann surface. One can say that all of the constructive work is done on the topological side, and that the topological results are realized in a function-theoretic manner with the help of a universal transfer principle, namely the Dirichlet principle. This is, in a sense, analogous to analytic geometry, where all the constructive work is carried out in the realm of numbers, and then the results are geometrically “realized” with the help of the transfer principle lodged in the coordinate concept.

All this is seen more perfectly in uniformization theory, which plays a central role in all of function theory. But at this point, I prefer to point to another application which is probably close to many of you. I have in mind enumerative geometry, which deals with the determination of the number of points of intersection, singularities, and so on, of algebraic relational structures, which was made into a general, but very poorly justified, system by Schubert and Zeuthen. Here, in the hands of Lefschetz and v.d. Waerden, topology achieved a decisive success in that it led to definitions of multiplicity valid without exception, as well as to laws likewise valid without exception. Of two curves on a two-sided surface one can cross the other at a point of intersection from left to right or from right to left. These points of intersection must enter every setup with opposite weights $+1$ and -1 . Then the total of the weights of the intersections (which can be positive or negative) is invariant under arbitrary continuous deformations of the curves; in fact, it remains unchanged if the curves are replaced by homologous curves. Hence it is possible to master this number through finite combinatorial means of topology and obtain transparent general formulas. Two algebraic curves are, actually, two closed Riemann surfaces embedded in a space of four real dimensions by means of an analytic mapping. But in algebraic geometry a point of intersection is counted with positive multiplicity, whereas in topology one takes into consideration the sense of the crossing. This being so, it is surprising that one can resolve the algebraic question by topological means. The explanation is that in the case of an analytic manifold, crossing always takes place with the same sense. If the two curves are represented in the x_1, x_2 -plane in the vicinity of their point of intersection by the functions $x_1 = x_1(s)$, $x_2 = x_2(s)$, and $x_1 = x_1^*(t)$, $x_2 = x_2^*(t)$, then the sense ± 1 with which the first curve intersects the second is given by the sign of the

$$\begin{vmatrix} \frac{dx_1}{ds} & \frac{dx_2}{ds} \\ \frac{dx_1^*}{dt} & \frac{dx_2^*}{dt} \end{vmatrix} = \frac{\partial(x_1, x_2)}{\partial(x, t)},$$

evaluated at the point of intersection. In the case of complex-algebraic “curves” this criterion always yields the value $+1$. Indeed, let z_1, z_2 be complex coordinates in the plane and let s and t be the respective complex parameters on the two “curves.” The real and imaginary parts of z_1 and z_2 play the role of real coordinates in the plane. In their place we can take $z_1, \bar{z}_1, z_2, \bar{z}_2$. But then the determinant whose sign determines the sense of the crossing is

$$\frac{\partial(z_1, \bar{z}_1, z_2, \bar{z}_2)}{\partial(s, \bar{s}, t, \bar{t})} = \frac{\partial(z_1, z_2)}{\partial(s, t)} \cdot \frac{\partial(\bar{z}_1, \bar{z}_2)}{\partial(\bar{s}, \bar{t})} = \left| \frac{\partial(z_1, z_2)}{\partial(s, t)} \right|^2,$$

and thus invariably positive. Note that the Hurwitz theory of correspondence between algebraic curves can likewise be reduced to a purely topological core.

On the side of abstract algebra, I will emphasize just one fundamental concept, namely the concept of an ideal. If we use the algebraic method, then an algebraic manifold is given in 3-dimensional space with complex cartesian coordinates x, y, z by means of a number of simultaneous equations

$$f_1(x, y, z) = 0, \dots, f_n(x, y, z) = 0.$$

The f_i are polynomials. In the case of a curve it is not at all true that two equations suffice. Not only do the polynomials f_i vanish at points of the manifold but also every polynomial f of the form

$$(**) \quad f = A_1 f_1 + \dots + A_n f_n \quad (A_i \text{ are polynomials}).$$

Such polynomials f form an “ideal” in the ring of polynomials. Dedekind defined an ideal in a given ring as a system of ring elements closed under addition and subtraction as well as under multiplication by ring elements. This concept is not too broad for our purposes. The reason is that, according to the Hilbert basis theorem, every ideal in the polynomial ring has a finite basis; there are finitely many polynomials f_1, \dots, f_n in the ideal such that every polynomial in the ideal can be written in the form $(**)$. Hence the study of algebraic manifolds reduces to the study of ideals. On an algebraic surface there are points and algebraic curves. The latter are represented by ideals that are divisors of the ideal under consideration. The fundamental theorem of M. Noether deals with ideals whose manifold of zeros consists of finitely many points, and makes membership of a polynomial in such an ideal dependent on its behavior at these points. This theorem follows readily from the decomposition of an ideal into prime ideals. The investigations of E. Noether show that the concept of an ideal, first introduced by Dedekind in the theory of algebraic number fields, runs through all of algebra and arithmetic like Ariadne’s thread. v.d. Waerden was able to justify the enumerative calculus by means of the algebraic resources of ideal theory.

If one operates in an arbitrary abstract number field rather than in the continuum of complex numbers, then the fundamental theorem of algebra, which asserts that every complex polynomial in one variable can be [uniquely] decomposed into linear factors, need not hold. Hence the general prescription in algebraic work: See if a proof makes use of the fundamental theorem or not. In

every algebraic theory there is a more elementary part that is independent of the fundamental theorem, and therefore valid in every field, and a more advanced part for which the fundamental theorem is indispensable. The latter part calls for the algebraic closure of the field. In most cases the fundamental theorem marks a crucial split; its use should be avoided as long as possible. To establish theorems that hold in an arbitrary field it is often useful to embed the given field in a larger field. In particular, it is possible to embed any field in an algebraically closed field. A well-known example is the proof of the fact that a real polynomial can be decomposed over the reals into linear and quadratic factors. To prove this, we adjoin i to the reals and thus embed the latter in the algebraically closed field of complex numbers. This procedure has an analogue in topology which is used in the study and characterization of manifolds; in the case of a surface, this analogue consists in the use of its covering surfaces.

At the center of today's interest is noncommutative algebra in which one does not insist on the commutativity of multiplication. Its rise is dictated by concrete needs of mathematics. Composition of operations is a kind of noncommutative operation. Here is a specific example. We consider the symmetry properties of functions $f(x_1, x_2, \dots, x_n)$ of a number of arguments. The latter can be subjected to an arbitrary permutation s . A symmetry property is expressed in one or more equations of the form

$$\sum_s a(s) \cdot sf = 0.$$

Here $a(s)$ stands for the numerical coefficients associated with the permutation. These coefficients belong to a given field K . $\sum_s a(s) \cdot s$ is a "symmetry operator." These operators can be multiplied by numbers, added and multiplied, that is, applied in succession. The result of the latter operation depends on the order of the "factors." Since all formal rules of computation hold for addition and multiplication of symmetry operators, they form a "noncommutative ring" (hypercomplex number system). The dominant role of the concept of an ideal persists in the noncommutative realm. In recent years, the study of groups and their representations by linear substitutions has been almost completely absorbed by the theory of noncommutative rings. Our example shows how the multiplicative group of $n!$ permutations s is extended to the associated ring of magnitudes $\sum_s a(s) \cdot s$ that admit, in addition to multiplication, addition and multiplication by numbers. Quantum physics has given noncommutative algebra a powerful boost.

Unfortunately, I cannot here produce an example of the art of building an abstract-algebraic theory. It consists in setting up the right general concepts, such as fields, ideals, and so on, in decomposing an assertion to be proved into steps (for example, an assertion " A implies B ," or $A \rightarrow B$, may be decomposed into steps $A \rightarrow C, C \rightarrow D, D \rightarrow B$), and in the appropriate generalization of these partial assertions in terms of general concepts. Once the main assertion has been subdivided in this way and the inessential elements have been set aside, the proofs of the individual steps do not, as a rule, present serious difficulties.

Whenever applicable, the topological method appears, thus far, to be more effective than the algebraic one. Abstract algebra has not yet produced successes comparable to the successes of the topological method in the hands of Riemann. Nor has anyone reached by an algebraic route the peak of uniformization scaled topologically by Klein, Poincaré and Koebe. Here are questions to be answered in the future. But I do not want to conceal from you the growing feeling among mathematicians that the fruitfulness of the abstracting method is close to exhaus-

tion. It is a fact that beautiful general concepts do not drop out of the sky. The truth is that, to begin with, there are definite concrete problems, with all their undivided complexity, and these must be conquered by individuals relying on brute force. Only then come the axiomatizers and conclude that instead of straining to break in the door and bloodying one's hands one should have first constructed a magic key of such and such shape and then the door would have opened quietly, as if by itself. But they can construct the key only because the successful breakthrough enables them to study the lock front and back, from the outside and from the inside. Before we can generalize, formalize and axiomatize there must be mathematical substance. I think that the mathematical substance on which we have practiced formalization in the last few decades is near exhaustion and I predict that the next generation will face in mathematics a tough time.

[The sole purpose of this lecture was to give the audience a feeling for the intellectual atmosphere in which a substantial part of modern mathematical research is carried out. For those who wish to penetrate more deeply I give a few bibliographical suggestions. The true pioneers of abstract axiomatic algebra are Dedekind and Kronecker. In our own time, this orientation has been decisively advanced by Steinitz, by E. Noether and her school, and by E. Artin. The first great advance in topology came in the middle of the 19th century and was due to Riemann's function theory. The more recent developments are linked primarily to a few works of H. Poincaré devoted to analysis situs (1895–1904). I mention the following books:

REFERENCES

- 1.. *On algebra*: Steinitz, *Algebraic Theory of Fields*, appeared first in *Crelles Journal* in 1910. It was issued as a paperback by R. Baer and H. Hasse and published by Verlag W. de Gruyter, 1930.
H. Hasse, *Higher algebra I, II*. Sammlung Götschen 1926/27.
B. v.d. Waerden, *Modern algebra I, II*. Springer 1930/31.
2. *On topology*: H. Weyl, *The Idea of a Riemann Surface*, second ed. Teubner 1923.
O. Veblen, *Analysis Situs*, second ed., and S. Lefschetz, *Topology*. Both of these books are in the series Colloquium Publications of the American Mathematical Society, New York 1931 and 1930 respectively.
3. Volume I of F. Klein, *History of Mathematics in the 19th Century*, Springer 1926.

Gauss once said, "Mathematics is the queen of the sciences and number theory the queen of mathematics." If this is true we may add that the disquisitions is the Magna Charter of number theory.

—M. Cantor
Allgemeine Deutsche Biographie, Bd. 8. 1878 p. 435.

PROBLEMS AND SOLUTIONS

Edited by:

Richard T. Bumby, Fred Kochman and Douglas B. West

Proposed problems should be sent to the MONTHLY PROBLEMS address given on the inside front cover. Please include solutions and relevant references. Three copies of all items needed to evaluate the problem should be sent.

Solutions of published problems should arrive at the MONTHLY PROBLEMS address given on the inside front cover before February 29, 1996. If possible, solutions should be typed with double spacing. Two copies suffice. Several solutions may be mailed together, but they should be on separate sheets of paper. The problem number and the solver's name and mailing address should appear on each solution. A mailing label should be included if an acknowledgment is desired.

The published solution is likely to be based on a solution that is complete and correct. Additional information, such as references to other appearances of the problem or its solution, is also welcome.

An asterisk () after the number of a problem, or part of a problem, indicates that no solution is currently available.*

PROBLEMS

10466. *Proposed by E. Sparre Andersen & Mogens Esrom Larsen, Københavns Universitet, København, Denmark.*

For $x \in \mathbb{C}$ and $n \in \mathbb{N}$, prove the following identities between polynomials.

(a) $(-4)^n \sum_{j=0}^n \binom{x + 1/2}{j} \binom{n-1-x}{2n-j} = \binom{2n}{n} \sum_{j=0}^n \binom{x+j}{2j} \binom{x-j}{2n-2j}$. For all $m \in \mathbb{N}$,

with $0 \leq m \leq 2n$, generalize (a) to

(b) $(-4)^n \sum_{j=0}^n \binom{x + 1/2}{j} \binom{n-1-x}{2n-j} = \binom{2n}{n} \sum_{j=-\lfloor m/2 \rfloor}^{n-\lfloor m/2 \rfloor} \binom{x+j}{2j+m} \binom{x-j}{2n-m-2j}$.

10467. *Proposed by Joseph E. Higgins, Cadence Design Systems, Inc., San José, CA.*

It seems geometrically evident that in a normed space X , the operator $\phi: X \setminus \{0\} \rightarrow X$ defined by $\phi(x) = x / |x|$ would satisfy the Lipschitz inequality $|\phi(x) - \phi(y)| \leq |x - y|$ whenever $|x| \geq 1$ and $|y| \geq 1$. Prove, or give a counterexample.

10468. Proposed by James E. Baumgartner & Benjamin J. Tilly (student), Dartmouth College, Hanover, NH.

Let F be a field and let $F^{\mathbb{N}}$ be the F -vector space of all functions from the nonnegative integers into F . What is the dimension of $F^{\mathbb{N}}$?

10469. Proposed by Jean Anglesio, Garches, France.

Let P be a point in the interior of the triangle ABC and let the lines AP , BP , CP meet the sides BC , CA , AB respectively at the points D , E , F . Let the circles on diameters BC and AD intersect at points a , a' ; the circles on diameters CA and BE intersect at points b , b' ; and the circles on diameters AB and CF intersect at points c , c' . Show that a , a' , b , b' , c , c' lie on a circle.

10470. Proposed by Donald E. Knuth, Stanford University, Stanford, CA.

Call a matrix (a_{ij}) *special* if its entries satisfy

$$a_{ij} = \begin{cases} 0, & \text{if } j > i + 1; \\ -1, & \text{if } j = i + 1; \\ 0 \text{ or } 1, & \text{if } j \leq i. \end{cases}$$

Call a special matrix *minimal* if its determinant is zero, but the determinant becomes nonzero when any element on or below the diagonal is changed from 0 to 1. For example,

$$\begin{array}{cccc} 1 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 1 & 1 & 1 & -1 \\ 0 & 0 & 1 & 0 \end{array}$$

is a minimal special matrix.

- (a) Prove that there are exactly 2^{n-1} minimal special matrices of size n by n .
- (b) What is the largest number of zeros that a minimal special matrix can contain?

10471. Proposed by Stephen Semmes & Richard Stong, Rice University, Houston, TX.

Suppose V is a (possibly infinite dimensional) vector space over \mathbb{C} and P_1, P_2, \dots, P_n are projections on V . For which n is it the case that if $P_1 + P_2 + \dots + P_n = 0$, then all the P_i are zero?

10472. Proposed by Edward Kitchen, Santa Monica, CA.

Let $P_0 P_1 P_2 P_3 P_4$ be a convex pentagon that is affinely equivalent to a regular pentagon. Let L_j be the center of a rotation through $+\pi/5$ radians taking P_{j+2} to P_{j-2} (all subscripts modulo 5). Show that P_j is the center of a rotation through $-\pi/5$ radians taking L_{j-1} to L_{j+1} .

NOTES

(10471) A *projection on V* is a linear function $P : V \rightarrow V$ such that $P^2 = P$. (10472) In chapter IX of *A Survey of Modern Algebra* by Birkhoff and MacLane, the *affine group* in a vector space is constructed from translations and nonsingular linear transformations. If one figure can be taken to another by an element of this group, the figures are said to be *affinely equivalent*.

SOLUTIONS

The Inverse of a Block Matrix

10186[1992, 60]. *Proposed by Lawrence A. Harris, University of Kentucky, Lexington, KY.*

Suppose A , B , and C are matrices of size m by n , m by m , and n by m , respectively, and suppose that CA is equal to the n by n identity matrix. Give a necessary and sufficient condition for the block matrix

$$M = \begin{pmatrix} A & B \\ 0 & C \end{pmatrix}$$

to be invertible and find an expression for M^{-1} when this condition holds.

Solution by University of Wyoming Problem Circle, University of Wyoming, Laramie, WY. Suppose that A , B and C satisfy the conditions of the problem. Define the m by m matrices $P := AC$ and $Q := I - AC$. Then the following statements are equivalent:

- (1) M is invertible;
- (2) $\mathcal{N}(QB) \cap \mathcal{N}(C) = \{0\}$;
- (3) the m by m matrix $P + QBQ$ is invertible.

(In (2), $\mathcal{N}(X)$ denotes the nullspace of the matrix X .) When condition (3) holds, introduce $N := (P + QBQ)^{-1}Q$ and $V := A - NBA$. Then one has

$$M^{-1} = \begin{pmatrix} C - CBN & -CBV \\ N & V \end{pmatrix}.$$

The following useful identities follow from $CA = I$: $CP = C$, $CQ = 0$, $PA = A$, $QA = 0$, $P + Q = I$, $P^2 = P$, $Q^2 = Q$, $PQ = 0 = QP$.

Proof that (1) \Rightarrow (2). Let $y \in \mathcal{N}(QB) \cap \mathcal{N}(C)$. Then $\begin{pmatrix} -CB y \\ y \end{pmatrix} \in \mathcal{N}(M)$ since

$$\begin{pmatrix} A & B \\ 0 & C \end{pmatrix} \begin{pmatrix} -CB y \\ y \end{pmatrix} = \begin{pmatrix} A(-CB y) + By \\ Cy \end{pmatrix} = \begin{pmatrix} QB y \\ Cy \end{pmatrix}.$$

Assuming (1), M is invertible, so $\mathcal{N}(M) = \{0\}$, giving $y = 0$.

Proof that (2) \Rightarrow (3). Let $y \in \mathcal{N}(P + QBQ)$. Then $0 = C(P + QBQ)y = CPy + CQBQy = CPy = Cy$. Hence $Py = ACy = 0$ and $Qy = y$. Consequently, $0 = (P + QBQ)y = QB y$. Thus, $\mathcal{N}(P + QBQ) \subseteq \mathcal{N}(QB) \cap \mathcal{N}(C)$. Assuming (2), this is $\{0\}$ and $P + QBQ$ is one-to-one, hence invertible.

Proof that (3) \Rightarrow (1). Assuming (3), we have the matrices N and V introduced above. Since $(P + QBQ)P = P^2 = P$, we have $(P + QBQ)^{-1}P = P$. Thus $NBQ = (P + QBQ)^{-1}QBQ = (P + QBQ)^{-1}((P + QBQ) - P) = I - P = Q$. Now, $NB + VC = NB + AC - NBAC = NB(I - AC) + AC = NBQ + P = Q + P = I$. Also $NA = 0$. To complete the proof, we verify the formula for M^{-1} .

$$\begin{aligned} \begin{pmatrix} C - CBN & -CBV \\ N & V \end{pmatrix} \begin{pmatrix} A & B \\ 0 & C \end{pmatrix} &= \begin{pmatrix} CA - CBN A & CB - CBN B - CBVC \\ NA & NB + VC \end{pmatrix} \\ &= \begin{pmatrix} I & CB(I - NB - VC) \\ 0 & NB + VC \end{pmatrix} \\ &= \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}. \end{aligned}$$

Editorial comment. Several solvers noted that the condition $CA = I$ implies that $n \leq m$. The proposer's solution verified directly that one can replace the given B by the identity matrix to obtain an invertible matrix M_1 , and M is invertible if and only if $M_1^{-1}M$ is. The latter matrix is a block upper triangular matrix with diagonal blocks of I and $P + QB$ and thus is invertible (with the inverse given by a standard formula) if and only if $P + QB$ is invertible. While the selected solution uses finite dimensionality, since invertibility is deduced from having a trivial nullspace, the proposer's approach may be easily extended to infinite dimensional spaces. Indeed, the problem was inspired by research at this level of generality (see Lawrence A. Harris, "Linear fractional transformations of circular domains in operator spaces", *Indiana Univ. Math. J.* 41 (1992), 125–147).

Solved also by Y. Ikeda, O. Krafft (Germany), J. H. Lindsey II, A. Nijenhuis, J. H. Steelman, O. Wyler, Westmont College Problem Solving Group, and the proposer. Seven incorrect or incomplete solutions were received.

A Semi-unfriendly Identity

10206[1992, 266]. *Proposed by David M. Bloom, Brooklyn College of CUNY, Brooklyn, NY.*

If m and k are positive integers, prove that

$$\sum_r \binom{r}{k-r} \binom{m}{r} = \sum_j \binom{\lfloor j/2 \rfloor}{k-j} \binom{m-k+\lfloor 3j/2 \rfloor}{j}.$$

Solution by Robin J. Chapman, University of Exeter, Exeter, U. K.. We analyze the generating functions of the two sides of the proposed identity as formal power series. For the left side:

$$\begin{aligned} L(x, y) &= \sum_{m,k,r} \binom{r}{k-r} \binom{m}{r} x^m y^k \\ &= \sum_{k,r} \binom{r}{k-r} \frac{x^r y^k}{(1-x)^{r+1}} \\ &= \sum_r \frac{x^r y^r (1+y)^r}{(1-x)^{r+1}} \\ &= \frac{1}{1-x} \sum_r \left[\frac{xy(1+y)}{1-x} \right]^r \\ &= \frac{1}{(1-x) - xy(1+y)} \\ &= \frac{1}{1-x(1+y+y^2)}. \end{aligned}$$

For the right side:

$$\begin{aligned} R(x, y) &= \sum_{m,k,j} \binom{\lfloor j/2 \rfloor}{k-j} \binom{m-k+\lfloor 3j/2 \rfloor}{j} x^m y^k \\ &= \sum_{j,k} \binom{\lfloor j/2 \rfloor}{k-j} \frac{(xy)^k}{x^{\lfloor j/2 \rfloor} (1-x)^{j+1}} \\ &= \sum_j \frac{(1+xy)^{\lfloor j/2 \rfloor} (xy)^j}{x^{\lfloor j/2 \rfloor} (1-x)^{j+1}} \end{aligned}$$

$$\begin{aligned}
&= \sum_i \left[\frac{(1+xy)^i (xy)^{2i}}{x^i (1-x)^{2i+1}} + \frac{(1+xy)^i (xy)^{2i+1}}{x^i (1-x)^{2i+2}} \right] \\
&= \frac{1-x+xy}{(1-x)^2} \sum_i \left[\frac{(1+xy)xy^2}{(1-x)^2} \right]^i \\
&= \frac{1-x+xy}{(1-x)^2 - (1+xy)xy^2} \\
&= \frac{1-x+xy}{1-2x+x^2-xy^2-x^2y^3} \\
&= \frac{1}{1-x(1+y+y^2)}.
\end{aligned}$$

This shows $L(x, y) = R(x, y)$, proving the proposed identity.

Editorial comment. The technique in this solution is called the *Snake Oil* method in H. S. Wilf, *Generatingfunctionology*, Academic Press, 1990. Other solvers cited D. Zeilberger, “The method of creative telescoping”, *J. Symbolic Computation* 11 (1991), 195–204 for algorithms showing that both sides of the given identity satisfy the recurrence

$$\begin{aligned}
&-3(m+1)(m+2)S(m, k) + (m+2)(-3k+7m+9)S(m+1, k) \\
&\quad + (k-2m-3)(-k+2m+4)S(m+2, k) = 0.
\end{aligned}$$

In one solution, the sum on the right was split into terms with odd j and terms with even j , and recurrences found for these parts separately. This gave a higher order recurrence for this expression, but the proof could still be completed by examining a finite number of values of m . The proposer observed that expression on the left is the number of k element subsets of $\{1, 2, \dots, m+k-1\}$ that contain no three consecutive integers. None of the proofs found a combinatorial identification of the expression on the right with this quantity. Another aspect of these *semi-unfriendly* subsets appeared in problem 10343 [93, 874].

Solved also by S. B. Ekhad, I. Nemes (Austria), and J. H. Steelman.

Fields Closed under n th Roots

10274[1993, 75]. *Proposed by Robert E. Byerly, Texas Tech University, Lubbock, TX.*

For odd integers n , let E_n be the smallest subfield of the real numbers closed under the function $x \rightarrow \sqrt[n]{x}$.

(a) If $f(x) \in \mathbb{Z}[x]$ is irreducible in $\mathbb{Z}[x]$, show that $f(x)$ has at most one root in E_3 .

(b) Are there any such fields E_n and \mathbb{Z} -irreducible polynomials $f(x)$ for which $f(x)$ has more than one root in E_n ?

Solution by Burt Fein and Robby Robson, Oregon State University, Corvallis, OR.

Let n be an odd number and let $f(x) \in \mathbb{Z}[x]$ be irreducible. We show that $f(x)$ cannot have two distinct roots in E_n by proving two lemmas:

Lemma 1. *There are no non-trivial monomorphisms of the field E_n into itself.*

Lemma 2. *Let $K \subseteq E_n$ be a subfield and suppose that $\sigma: K \rightarrow E_n$ is a monomorphism. Then σ extends to a monomorphism defined for all of E_n .*

These two lemmas imply the result, for if α and β are roots of $f(x)$ in E_n , then the monomorphism $\sigma: \mathbb{Q}(\alpha) \rightarrow E_n$ having $\sigma(\alpha) = \beta$ extends to all of E_n by Lemma 2, and this must be the identity by Lemma 1. Hence $\alpha = \beta$.

Proof of Lemma 1. Suppose $\sigma: E_n \rightarrow E_n$ is a non-trivial monomorphism. Let L be the subfield of E_n fixed by σ . Since σ is non-trivial, L is a proper subfield. By the definition

of E_n , there is an $x \in L$ such that $\sqrt[n]{x} \in E_n - L$. Since $\sigma(\sqrt[n]{x}) \neq \sqrt[n]{x}$, E_n contains two real n th roots of x , which is impossible.

Proof of Lemma 2. After applying Zorn's Lemma to pairs (L, ϕ) consisting of subfields L of E_n containing K and monomorphisms $\phi: L \rightarrow E_n$ extending σ , we may assume that (K, σ) is a maximal pair. Our goal is to show that $K = E_n$.

If not, there is some $a \in K$ with $\sqrt[n]{a} \notin K$. The polynomial $x^n - a$ has only one real root and has odd degree. Thus, if we consider its factorization in $K[x]$ into irreducible polynomials

$$x^n - a = f_1(x)f_2(x) \cdots f_m(x),$$

exactly one of the f_i has odd degree. Assume f_1 has odd degree, and hence that $\sqrt[n]{a}$ is a root of $f_1(x)$. Applying σ to the factorization of $x^n - a$, we find that the element $\sqrt[n]{\sigma(a)}$, which is in E_n because of the way E_n is defined, is the unique real root of

$$x^n - \sigma(a) = f_1^\sigma(x)f_2^\sigma(x) \cdots f_m^\sigma(x).$$

Since only f_1^σ has odd degree, $\sqrt[n]{\sigma(a)}$ is a root of $f_1^\sigma(x)$. It is a well-known result (see, for example, Theorem 1.8 in T. W. Hungerford, *Algebra*, Springer-Verlag, 1974, p. 261) that this condition is sufficient for σ to be extendable to a monomorphism from $K(\sqrt[n]{a})$ into E_n sending $\sqrt[n]{a}$ to $\sqrt[n]{\sigma(a)}$. This contradicts the maximality of the pair (K, σ) and establishes the lemma.

Solved also by S. Ott, F. Richman, and the proposer.

An Unsettled Inequality

10337[1993, 798]. *Proposed by Horst Alzer, Waldbrohl, Germany.*

Let $n \geq 1$ be an integer. Let x_1, \dots, x_n be real numbers with $x_i \in (0, 1/2]$. Consider the statement

$$\prod_{i=1}^n \frac{x_i}{1-x_i} \leq \frac{\sum_{i=1}^n x_i^n}{\sum_{i=1}^n (1-x_i)^n}. \quad (\mathbf{F}_n)$$

- (a) Prove \mathbf{F}_n for $n \leq 3$.
- (b) Show that \mathbf{F}_n is false for $n \geq 6$.
- (c) What about \mathbf{F}_4 and \mathbf{F}_5 ?

Solution by Michael Vowe, Therwil, Switzerland.

Part (a):

\mathbf{F}_1 is trivial.

\mathbf{F}_2 is equivalent to $(1-x_1-x_2)(x_1-x_2)^2 \geq 0$, and thus true.

\mathbf{F}_3 can be rewritten as

$$\frac{1}{2} \sum_{i=1}^3 (x_i - x_{i+1})^2 \cdot \sum_{i=1}^3 (1-x_i)(1-x_{i+1}-x_{i+2}) \geq 0$$

(where $x_4 = x_1$ and $x_5 = x_2$) and thus true.

Part (b): Put $x_1 = x_2 = \dots = x_{n-1} = 1/2$ and $x_n = 1/5$. Then \mathbf{F}_n is false if

$$\frac{1}{4} > \frac{(n-1)2^{-n} + (0.2)^n}{(n-1)2^{-n} + (0.8)^n}.$$

This is equivalent to $4^n > 3(n-1)(2.5)^n + 4$, which holds for $n \geq 6$.

Editorial comment. Thomas L. McCoy investigated general properties of the statements \mathbf{F}_n using expressions in terms of elementary symmetric polynomials and power sums in the x_i . These studies allowed easy discovery of the equivalent form of \mathbf{F}_3 given in the

selected solution. There were some applications to F_4 and F_5 , but the truth or falsity of these statements remains unsettled. The proposer indicated that computer experiments by F. Bullock failed to locate any counterexamples to these statements.

Solved also by T. L. McCoy, H.-J. Seiffert (Germany), GCHQ Problem Solving Group (U. K.), and the proposer. One incorrect solution was received.

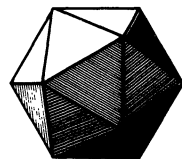
Collaborating editors: *David F. Appleyard, Paul T. Bateman, Duane M. Broline, Barry W. Brunson, Frank S. Cater, Gulbank D. Chakerian, Underwood Dudley, Gerald A. Edgar, Michael A. Filaseta, Ira M. Gessel, Richard A. Gibbs, Jerrold R. Griggs, Douglas A. Hensley, John R. Isbell, Mourad E. H. Ismail, Murray Klamkin, Daniel J. Kleitman, Frederick W. Luttman, Frank B. Miles, Richard Pfiefer, Stephen L. Portnoy, J. O. Shallit, John Henry Steelman, Kenneth B. Stolarsky, David E. Tepper, Douglas B. Tyler, Daniel Ullman, and William E. Watkins.*

THE COVER

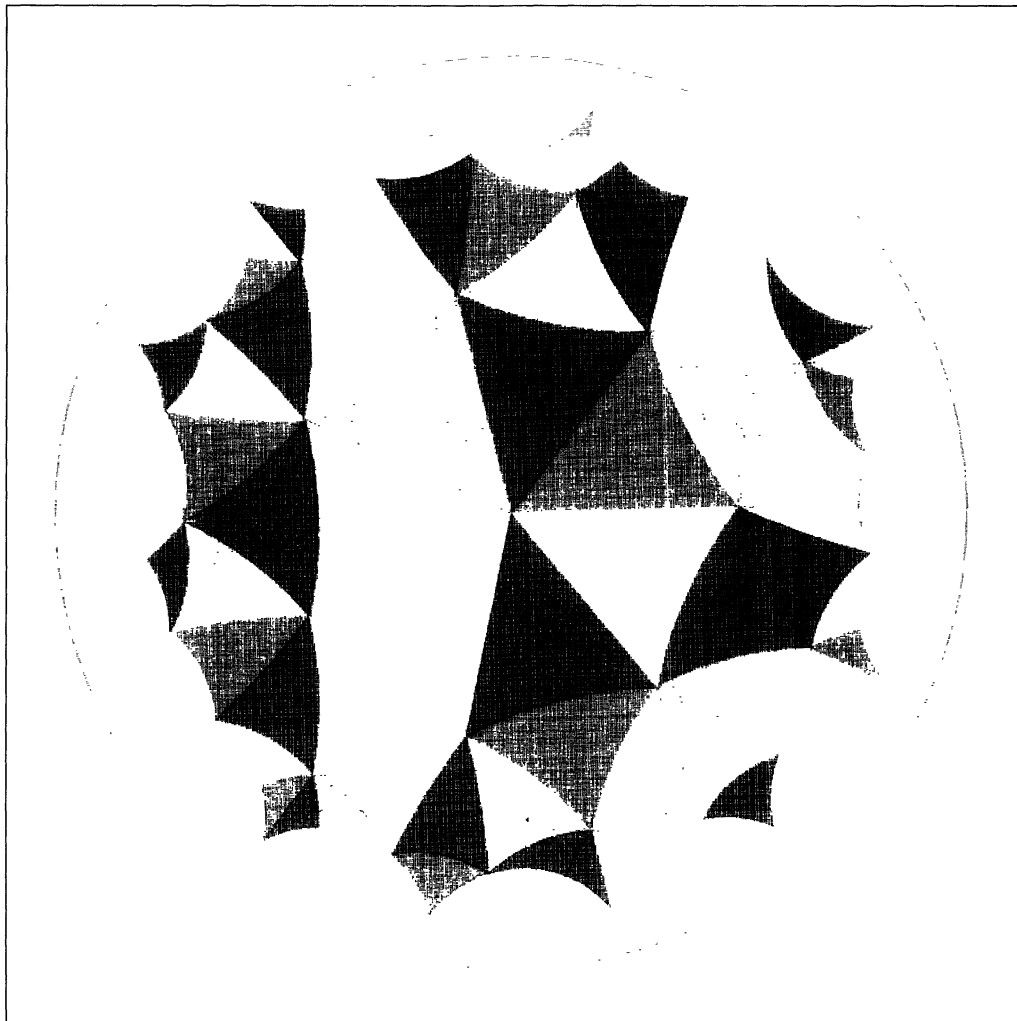
David Fowler would like to hear from you if you can recognize the function which is plotted on the cover, or if you have ever seen such a plot before. The answer will be in an article by him in the January 1996 issue. Contact him before January with your answers, and he will report on the response. So far, he has shown it to hundreds of people; nobody has seen the plot before, and only two people have recognized the underlying function without considerable help. Bits of the function, he adds, have been known to mathematicians for more than 900 years.

*David Fowler
Mathematics Institute
University of Warwick
Coventry CV4 7AL
ENGLAND
dhf@maths.warwick.ac.uk
Fax +44-1203-523548*

The American Mathematical Monthly



Volume 102, Number 8 / OCTOBER 1995



Hyperbolic Plane Colorings
(See page 706)

AN OFFICIAL PUBLICATION OF THE MATHEMATICAL ASSOCIATION OF AMERICA

NOTICE TO AUTHORS

The *Monthly* publishes articles, notes, and other features about mathematics and the profession. The readership of the *Monthly* is intended to include everybody who is mathematically inclined, including of course professional mathematicians and students of mathematics at all collegiate levels. While no single article or feature is likely to appeal to everyone, material should interest and be accessible to a large number of readers. This is the most important criterion for acceptance.

Articles may be expositions of old results or presentations of new ones. They may concern all of mathematics or one small area, a broad development or a single application, historical reminiscences or one important event. While some articles may contain the author's new research, the novelty of material and generality of the results is far less important than the clarity of exposition and general interest. Discussing one illuminating case of a well known result is far better than providing all the details of an obscure but new proposition. Articles in the *Monthly* are supposed to inform and to entertain; they are meant to be read rather than archived.

Notes are short and possibly informal articles. A note may concern a clever new proof of an old theorem, a novel way to present tired material, or a lively discussion of a philosophical (but still mathematical) issue. Also, any topic is suitable, so long as it is related to mathematics. Because a note is short, the first few sentences are the most important part: They should explain the purpose and invite the reader in. Photographs or diagrams often will attract the reader's attention.

All articles and notes should be sent to the editor:

ROGER HORN
1515 Mineral Square, Room 142
University of Utah
Salt Lake City, UT 84112

Please send 3 copies, typewritten on only one side of the paper. Illustrations should be carefully drawn on separate sheets of paper in black ink; the original should be without lettering and two copies should have appropriate captions and lettering indicated.

Proposed problems or solutions should be sent to:

RICHARD BUMBY,
P.O. Box 10971
New Brunswick, NJ 08906-0971.

Please send 2 copies of all material, typewritten if possible.

Letters to the Editor, both for publication and for private reading, should be sent to the Editor at the address given above. Comments, including criticisms, are welcome, as are all suggestions for making the *Monthly* a lively, entertaining, and informative journal.

EDITOR:

JOHN H. EWING

ASSOCIATE EDITORS:

PETER BORWEIN	FRED KOCHMAN
RICHARD BUMBY	CATHERINE MCGEOCH
DENNIS DETURCK	RICHARD NOWAKOWSKI
UNDERWOOD DUDLEY	ARNOLD OSTEBEE
JOHN DUNCAN	LEE RUBEL
JOAN FERRINI-MUNDY	ABE SHENITZER
JOSEPH GALLIAN	LYNN STEEN
STEVEN GALOVICH	STAN WAGON
RICHARD GUY	DOUGLAS WEST
DARRELL HAILE	HERBERT WILF
PAUL HALMOS	SANDY ZABELL
JOAN HUTCHINSON	PAUL ZORN

EDITORIAL ASSISTANT:

MISTY CUMMINGS

STAFF ARTIST:

MIKE CAGLE

QUOTE MASTER:

MARK WOODARD

Reprint permission:

MARCIA P. SWARD, Executive Director

Advertising Correspondence:

Ms. ELAINE PEDREIRA, Advertising Manager

Subscription correspondence, change of address, and other inquiries:

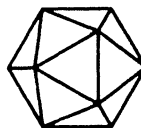
Membership / Subscriptions Department

All at the address:

The Mathematical Association of America
1529 Eighteenth Street, N.W.
Washington, DC 20036.

Microfilm Editions: University Microfilms International, Serial Bid coordinator, 300 North Zeeb Road, Ann Arbor, MI 48106.

The AMERICAN MATHEMATICAL MONTHLY (ISSN 0002-9890) is published monthly except bimonthly June-July and August-September by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, DC 20036 and Montpelier, VT. Copyrighted by the Mathematical Association of America (Incorporated), 1995, including rights to this journal issue as a whole and, except where otherwise noted, rights to each individual contribution. General permission is granted to Institutional Members of the MAA for noncommercial reproduction in limited quantities of individual articles (in whole or in part) provided a complete reference is made to the source. Second class postage paid at Washington, DC, and additional mailing offices. **Postmaster:** Send address changes to the American Mathematical Monthly, Membership / Subscription Department, MAA, 1529 Eighteenth Street, N.W., Washington, DC, 20036-1385.



Contents

ARTICLES

**The Fifty-Fifth William Lowell Putnam Mathematical Competition /
LEONARD F. KLOSINSKI, GERALD L. ALEXANDERSON,
AND LOREN C. LARSON 675**

**Mathematics: Questions and Answers / BENO ECKMANN
(translated by PETER HILTON) 685**

**Teaching Math More Effectively, Through Computational Proofs /
DAVID GRIES AND FRED B. SCHNEIDER 691**

**Some Problems Concerning Recurrence Sequences / G. MYERSON
AND A. J. VAN DER POORTEN 698**

**A Hyperbolic Plane Coloring and the Simple Group of Order 168 /
DANA MACKENZIE 706**

**Cosine Products, Fourier Transforms, and Random Sums /
KENT E. MORRISON 716**

How to Add Fast—on Average / GEZA SCHAY 725

FEATURES

COMMENTS 674

NOTES

**Fibonacci-like Sequences and Greatest Common Divisors /
H. R. MORTON 731**

**Generating Symmetric Groups / I. M. ISAACS AND
THILO ZIESCHANG 734**

**On the Arithmetic-Geometric Mean Inequality /
LUTZ G. LUCHT 739**

UNSOLVED PROBLEMS

**Three Open Problems in Functional Equations /
P. K. SAHOO 741**

THE AUTHORS 743

PROBLEMS AND SOLUTIONS 745

REVIEWS

***Computability.* By Douglas S. Bridges /**

YIANNIS N. MOSCHOVAKIS 752

***Knot Theory.* By Charles Livingston / JOAN S. BIRMAN 755**

TELEGRAPHIC REVIEWS 758

The Fifty-Fifth William Lowell Putnam Mathematical Competition

**Leonard F. Klosinski, Gerald L. Alexanderson
and Loren C. Larson**

The following results of the fifty-fifth William Lowell Putnam Mathematical Competition, held on December 3, 1994, have been determined in accordance with the governing regulations. This annual contest is supported by the William Lowell Putnam Prize Fund for the Promotion of Scholarship, left by Mrs. Putnam in memory of her husband, and is held under the auspices of the Mathematical Association of America.

The first prize, \$7,500, was awarded to the Department of Mathematics at Harvard University. The members of the winning team were: Kiran S. Kedlaya, Lenhard L. Ng, and Dylan P. Thurston; each was awarded a prize of \$500.

The second prize, \$5,000, was awarded to the Department of Mathematics at Cornell University. The members of the winning team were Jeremy L. Bem, Robert D. Kleinberg, and Mark Krosky; each was awarded a prize of \$400.

The third prize, \$3,000, was awarded to the Department of Mathematics at the Massachusetts Institute of Technology. The members of the winning team were Henry L. Cohn, Adam W. Meyerson, and Thomas A. Weston; each was awarded a prize of \$300.

The fourth prize, \$2,000, was awarded to the Department of Mathematics at Princeton University. The members of the winning team were William R. Mann, Joël E. Rosenberg, and Michail Sunitzky; each was awarded a prize of \$200.

The fifth prize, \$1,000, was awarded to the Department of Mathematics of the University of Waterloo. The members of the winning team were Ian A. Goldberg, Peter L. Milley, and Kevin Purbhoo; each was awarded a prize of \$100.

The five highest ranking individual contestants, in alphabetical order, were Jeremy L. Bem, Cornell University; J. P. Grossman, University of Toronto; Kiran S. Kedlaya, Harvard University; William R. Mann, Princeton University; and Lenhard L. Ng, Harvard University. Each of these was designated a Putnam Fellow by the Mathematical Association of America and awarded a prize of \$1,000, by the Putnam Prize Fund.

The next five highest ranking contestants, in alphabetical order, were Soundararajan Kannan, University of Michigan, Ann Arbor; David L. Savitt, University of British Columbia; Daniel K. Schepler, Washington University, St. Louis; Noam M. Shazeer, Duke University; and Hong Zhou, Harvard University; each was awarded a prize of \$500.

The next six highest ranking contestants, in alphabetical order, were Alexandru D. Ionescu, Massachusetts Institute of Technology; Robert D. Kleinberg, Cornell University; Jacob A. Rasmussen, Princeton University; Andrew H. Schultz, Johns Hopkins University; Dylan P. Thurston, Harvard University; and Zhaohui Zhang, Yale University; each was awarded a prize of \$250.

The next nine highest ranking contestants, in alphabetical order, were Henry L. Cohn, Massachusetts Institute of Technology; Ian A. Goldberg, University of Waterloo; Adam Kalai, Harvard University; Serban M. Nacu, Harvard University; Joel E. Rosenberg, Princeton University; Mikhail V. Shubov, Texas Tech University; Jade P. Vinson, Washington University, St. Louis; Stephen S. Wang, Harvard University; and Jonathan L. Weinstein, Harvard University. Each was awarded a prize of \$100.

The following teams, named in alphabetical order, received honorable mention: University of Nebraska, Lincoln, with team members Scott Annin, Igor V. Pavlovsky, and Eric M. Smith; New York University, with team members Igor Berger, Yevgeniy Dodis, and Mikhail Kogan; University of Toronto, with team members J. P. Grossman, Edward Leung, and Naoki Sato; Washington University, St. Louis, with team members Ben Gum, Daniel K. Schepler, and Jade P. Vinson; and Yale University, with team members Gautam Chinta, Matthew Frank, and Zhaohui Zhang.

Honorable mention was achieved by the following thirty individuals named in alphabetical order: Jared E. Anderson, University of Victoria; Federico Ardila, Massachusetts Institute of Technology; Bradley S. Bart, University of Waterloo; Ruth A. Britto-Pacumio, Massachusetts Institute of Technology; Robert H. Cheng, University of British Columbia; Yevgeniy Dodis, New York University; Ron D. Dror, Rice University; Alex Heneveld, Princeton University; Randy W. Ho, University of Arizona; Jason A. Howald, Miami University; Sergey M. Ioffe, Massachusetts Institute of Technology; Dean W. Jens, University of Chicago; Joanna L. Karczmarek, Queen's University; Mikhail Kogan, New York University; Botond Kőszegi, Harvard University; Mark Krosky, Cornell University; Daniel T. Martin, Carleton College; Olexei Ivanovich Motrunich, University of Missouri, Columbia; Akira Negi, University of North Carolina, Chapel Hill; An T. Nguyen, University of Texas, Austin; Royce Y. Peng, Harvard University; Kevin Purbhoo, University of Waterloo; Lawrence P. Roberts, Washington University, St. Louis; NNaoki Sato, University of Toronto; Sam Spencer, Rice University; Jason M. Starr, University of California, Berkeley; Mark A. Van Raamsdonk, University of British Columbia; David R. Wasserman, University of California, San Diego; Thomas A. Weston, Massachusetts Institute of Technology; and Jeffrey S. Willson, University of Chicago.

The other individuals who achieved ranks among the top 107, in alphabetical order of their schools, were: Brown University, Andrew Brecher; California Institute of Technology, Wei-Hwa Huang, Roman Muchnik; California Polytechnic State University, San Luis Obispo, Robert B. Mathews; University of California, Santa Barbara, Aaron S. Cohen; Carleton College, Curtis Z. Mitchell; Case Western Reserve University, Neil A. Rubin; Colgate University, Jean-François R. Lafont; Dartmouth College, Yuan Shen; Duke University, Robert R. Schneck; Harvard University, Manjul Bhargava, Dean R. Chung, Joe B. Fendel, Sergey V. Levin, Paul Li, Harrison K. Tsai, Jiří J. L. Vaníček; Harvey Mudd College, Aaron F. Archer, Kan Yasuda; University of Illinois, Champaign-Urbana, Ivan Auramovic, Kwong Shing Lin; Massachusetts Institute of Technology, Adam W. Meyerson, Michael B. Schulz, Michael R. Tehranchi, Aleksey Zinger; McGill University, Jacob Eliosoff; University of Nebraska, Lincoln, Eric M. Smith; New York University, Igor Berger; University of North Carolina, Chapel Hill, Paul E. Rube; Northwestern University, Carol R. James; Princeton University, Paul J. Ellis, Michael J. Goldberg, Mark W. Lucianovic; Queen's University, Peter Gregory Zion; Reed College, Gerald D. Larson; Rice University, Ashley M.

Reiter; University of Saskatchewan, Trevor N. Green; University of the South, Qingshan Luo; Stanford University, Robert G. Au, Heyning A. Cheng, Loren L. Looger; Suffolk University, Anna V. Petrovskaya; Vanderbilt University, Jason D. Hughes; Washington University, St. Louis, Ian F. Pulizzotto, Erik N. Vee; University of Waterloo, Jason P. Bell, Jie J. Lou, Peter L. Milley, Lousindi R. Sabourin; Williams College, Jason R. Schweinsberg, Edward W. Welsh; and Yale University, Matthew Frank.

The Elizabeth Lowell Putnam Prize, named for the wife of William Lowell Putnam and to be “awarded periodically to a woman whose performance on the Competition has been deemed particularly meritorious,” is awarded this year to Ruth A. Britto-Pacumio of the Massachusetts Institute of Technology. The winner is awarded a prize of \$500.

There were 2,314 individual contestants from the 410 colleges and universities in Canada and the United States in the competition of December 3, 1994. Teams were entered by 284 institutions. The Questions Committee for the fifty-fifth competition consisted of Eugene M. Luks, University of Oregon, chair; Fan Chung, Bellcore; and Mark I. Krusemeyer, Carleton College; they composed the problems listed below and were most prominent among those suggesting solutions.

PROBLEMS

Problem A-1. Suppose that a sequence a_1, a_2, a_3, \dots satisfies $0 < a_n \leq a_{2n} + a_{2n+1}$ for all $n \geq 1$. Prove that the series $\sum_{n=1}^{\infty} a_n$ diverges.

Problem A-2. Let A be the area of the region in the first quadrant bounded by the line $y = \frac{1}{2}x$, the x -axis, and the ellipse $\frac{1}{9}x^2 + y^2 = 1$. Find the positive number m such that A is equal to the area of the region in the first quadrant bounded by the line $y = mx$, the y -axis, and the ellipse $\frac{1}{9}x^2 + y^2 = 1$.

Problem A-3. Show that if the points of an isosceles right triangle of side length 1 are each colored with one of four colors, then there must be two points of the same color which are at least a distance $2 - \sqrt{2}$ apart.

Problem A-4. Let A and B be 2×2 matrices with integer entries such that A , $A + B$, $A + 2B$, $A + 3B$, and $A + 4B$ are all invertible matrices whose inverses have integer entries. Show that $A + 5B$ is invertible and that its inverse has integer entries.

Problem A-5. Let $(r_n)_{n \geq 0}$ be a sequence of positive real numbers such that $\lim_{n \rightarrow \infty} r_n = 0$. Let S be the set of numbers representable as a sum

$$r_{i_1} + r_{i_2} + \cdots + r_{i_{1994}},$$

with $i_1 < i_2 < \cdots < i_{1994}$. Show that every nonempty interval (a, b) contains a nonempty subinterval (c, d) that does not intersect S .

Problem A-6. Let f_1, f_2, \dots, f_{10} be bijections of the set of integers such that for each integer n , there is some composition $f_{i_1} \circ f_{i_2} \circ \cdots \circ f_{i_m}$ of these functions (allowing repetitions) which maps 0 to n . Consider the set of 1024 functions

$$\mathcal{F} = \{f_1^{e_1} \circ f_2^{e_2} \circ \cdots \circ f_{10}^{e_{10}}\},$$

$e_i = 0$ or 1 for $1 \leq i \leq 10$. (f_i^0 is the identity function and $f_i^1 = f_i$.) Show that if A

is any nonempty finite set of integers, then at most 512 of the functions in \mathcal{F} map A to itself.

Problem B-1. Find all positive integers that are within 250 of exactly 15 perfect squares.

Problem B-2. For which real numbers c is there a straight line that intersects the curve

$$y = x^4 + 9x^3 + cx^2 + 9x + 4$$

in four distinct points?

Problem B-3. Find the set of all real numbers k with the following property: For any positive, differentiable function f that satisfies $f'(x) > f(x)$ for all x , there is some number N such that $f(x) > e^{kx}$ for all $x > N$.

Problem B-4. For $n \geq 1$, let d_n be the greatest common divisor of the entries of $A^n - I$, where

$$A = \begin{pmatrix} 3 & 2 \\ 4 & 3 \end{pmatrix} \quad \text{and} \quad I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Show that $\lim_{n \rightarrow \infty} d_n = \infty$.

Problem B-5. For any real number α , define the function $f_\alpha(x) = \lfloor \alpha x \rfloor$. Let n be a positive integer. Show that there exists an α such that for $1 \leq k \leq n$,

$$f_\alpha^k(n^2) = n^2 - k = f_\alpha^k(n^2).$$

Problem B-6. For any integer a , set

$$n_a = 101a - 100 \cdot 2^a.$$

Show that for $0 \leq a, b, c, d \leq 99$, $n_a + n_b \equiv n_c + n_d \pmod{10100}$ implies $\{a, b\} = \{c, d\}$.

SOLUTIONS. In the 12-tuples $(n_{10}, n_9, n_8, n_7, n_6, n_5, n_4, n_3, n_2, n_1, n_0, n_{-1})$ following each problem number below, n_i for $10 \geq i \geq 0$ is the number of students among the top 206 contestants achieving i points for the problem and n_{-1} is the number of those not submitting solutions.

A-1 (59, 59, 54, 21, 0, 0, 0, 0, 8, 0, 3, 2)

Solution. Let $b_1 = a_1$, $b_2 = a_2 + a_3$, $b_3 = a_4 + a_5 + a_6 + a_7$, and in general, $b_n = a_{2^{n-1}} + a_{2^{n-1}+1} + \cdots + a_{2^n-1}$. AN easy induction, using the condition $a_n \leq a_{2n} + a_{2n+1}$ shows that $b_n \leq b_{n+1}$ for all $n \geq 1$. Thus, for any positive integer t ,

$$\sum_{n=1}^{\infty} a_n > \sum_{n=1}^{2^t-1} a_n = \sum_{n=1}^t b_n \geq tb_1 = ta_1.$$

This shows that $\sum_{n=1}^{\infty} a_n$ diverges.

A-2 (169, 3, 2, 0, 0, 0, 0, 0, 1, 3, 22, 6)

Solution. The linear transformation given by $x_1 = \frac{1}{3}x$, $y_1 = y$ transforms the region R bounded by $y = \frac{1}{2}x$, the x -axis, and the ellipse $\frac{1}{9}x^2 + y^2 = 1$ into the

region R' bounded by $y_1 = \frac{3}{2}x_1$, the x_1 -axis, and the circle $x_1^2 + y_1^2 = 1$; it also transforms the region S bounded by $y = mx$, the y -axis, and $\frac{1}{9}x^2 + y^2 = 1$ into the region S' bounded by $y_1 = 3mx_1$, the y_1 -axis, and the circle. Since all areas are multiplied by the same (nonzero) factor under the transformation, R and S have the same area if and only if R' and S' have the same area. However, we can see by symmetry about the line $y_1 = x_1$ that this happens if and only if $3m = \frac{2}{3}$, that is, $m = \frac{2}{9}$.

A-3 (0, 10, 67, 0, 0, 0, 0, 30, 31, 40, 28)

Solution. Suppose the vertices of the isosceles right triangle are $(0, 0)$, $(1, 0)$, $(0, 1)$. Suppose the points of the triangle can be colored in four colors such that points of the same color are always less than a distance $2 - \sqrt{2}$ apart. Then the four points $(0, 1)$, $(0, \sqrt{2} - 1)$, $(\sqrt{2} - 1, 0)$, $(1, 0)$ must have different colors, say colors A, B, C, D respectively. The point $(0, 0)$ must be of color B or C . Without loss of generality, say $(0, 0)$ is of color B . Then the point $(\sqrt{2} - 1, 2 - \sqrt{2})$ is of distance at least $2 - \sqrt{2}$ to points of each of the four colors, and this is impossible.

A-4 (12, 17, 20, 0, 0, 0, 0, 15, 3, 43, 96)

Solution. A matrix C with integer entries has an inverse with integer entries if and only if $\det C = \pm 1$. Therefore, if we consider the function f defined by $f(x) = \det(A + xB)$, we know that the five values $f(0)$, $f(1)$, $f(2)$, $f(3)$, and $f(4)$ must all be 1 or -1 , so f takes on at least one of those values three or more times. However, $f(x)$ is a polynomial of degree ≤ 2 in x , and so f can only take on a value more than twice if f is constant. Thus $f(x)$ is one of the constants 1 and -1 ; in particular, $\det(A + 5B) = \pm 1$, so $A + 5B$ has an inverse with integer entries.

A-5 (20, 13, 4, 0, 0, 0, 0, 6, 2, 57, 104)

Solution 1. It suffices to show that any sequence in S contains a monotonically nonincreasing subsequence. For then, letting $(t_n)_{n \geq 0}$ be any strictly increasing sequence within (a, b) , some (in fact, all but a finite number) of the intersections $S \cap (t_n, t_{n+1})$ would have to be empty, otherwise one could form a strictly increasing sequence $(s_n)_{n \geq 0}$ by taking $s_n \in S \cap (t_n, t_{n+1})$.

Let $(s_n)_{n \geq 0}$ be a sequence in S . For $n = 0, 1, 2, \dots$ write

$$s_n = r_{f(n,1)} + r_{f(n,2)} + \dots + r_{f(n,1994)} \quad \text{with} \quad f(n,1) < f(n,2) < \dots < f(n,1994).$$

The sequence $(r_{f(n,1)})_{n \geq 0}$ has a monotonically nonincreasing subsequence (since $(r_n)_{n \geq 0}$ is a positive sequence converging to 0). Thus we may replace $(s_n)_{n \geq 0}$ by a subsequence for which $(r_{f(n,1)})_{n \geq 0}$ is monotonically nonincreasing. In a similar fashion, we pass to subsequences so that, successively, each of $(r_{f(n,2)})_{n \geq 0}$, $(r_{f(n,3)})_{n \geq 0}, \dots, (r_{f(n,1994)})_{n \geq 0}$ may be assumed to be monotonically nonincreasing. The resulting $(s_n)_{n \geq 0}$ is monotonically nonincreasing.

Solution 2. Let C be the set $\{r_n\}_{n \geq 0} \cup \{0\}$. Since C is compact, the set S' of numbers representable as a sum of 1994 elements of C is also compact (for example, it is a continuous image of C^{1994}). Clearly $S \subseteq S'$.

Let (a, b) be a nonempty open interval. Since S' is countable, $(a, b) \setminus S'$ is nonempty; it is open since S' is closed. Hence $(a, b) \setminus S'$ includes a nonempty open interval.

Comment: This proof generalizes to give the same conclusion for any convergent sequence $(r_n)_{n \geq 0}$.

A-6 (5, 8, 10, 0, 0, 0, 0, 7, 4, 34, 138)

Solution. Let A be a nonempty finite subset of the integers \mathbf{Z} . By the Pigeonhole Principle, any bijection of \mathbf{Z} which maps A to itself must be a bijection when restricted to A ; in particular, its inverse also maps A to itself. Note that not all the bijections f_1, f_2, \dots, f_{10} can map A to itself, for otherwise if $0 \in A$ we could not map 0 to any $n \notin A$ by a composition $f_{i_1} \circ f_{i_2} \circ \dots \circ f_{i_m}$, while if $0 \notin A$, we could not map 0 to any $n \in A$ by such a composition.

Let k be the smallest integer such that f_k does not map A to itself, and suppose that more than 512 of the functions \mathcal{F} map A to itself. We can write \mathcal{F} as a disjoint union of unordered pairs of functions such that two compositions $f_1^{e_1} \circ f_2^{e_2} \circ \dots \circ f_{10}^{e_{10}}$ and $f_1^{d_1} \circ f_2^{d_2} \circ \dots \circ f_{10}^{d_{10}}$ are in the same pair when they differ only in the k -th exponent; that is, when $e_i = d_i$ for $i \neq k$. By the Pigeonhole Principle, there is then at least one of these 512 pairs in which both functions map A to itself. Since all f_l with $l > k$ also map A to itself, we can use composition with the inverses of f_l , as needed, to conclude that for some e_1, \dots, e_{k-1} , $F_1 = f_1^{e_1} \circ f_2^{e_2} \circ \dots \circ f_{k-1}^{e_{k-1}}$ and $F_2 = f_1^{e_1} \circ f_2^{e_2} \circ \dots \circ f_{k-1}^{e_{k-1}} \circ f_k$ both map A to itself. But then $F_1^{-1} \circ F_2 = f_k$ also maps A to itself, a contradiction.

B-1 (45, 26, 57, 0, 0, 0, 0, 42, 28, 6, 2)

Solution. Answer: $\{N \mid 315 \leq N \leq 325 \text{ or } 332 \leq N \leq 350\}$.

Assume $N > 0$ is within 250 of the 15 squares $m^2, (m+1)^2, \dots, (m+14)^2$, where we can take $m \geq 0$. In fact, m will then be positive, otherwise N would be within 250 of the additional square 225. We have the necessary and sufficient conditions

$$(m+14)^2 \leq N+250 \leq (m+15)^2 - 1,$$

$$(m-1)^2 + 1 \leq N-250 \leq m^2.$$

Subtracting (reversing inequalities in the second line), we get

$$28m + 196 \leq 500 \leq 32m + 222,$$

which implies $m = 9$ or 10 .

If $m = 9$,

$$23^2 \leq N+250 \leq 24^2 - 1,$$

$$8^2 + 1 \leq N-250 \leq 9^2,$$

or $315 \leq N \leq 325$.

If $m = 10$,

$$24^2 \leq N+250 \leq 25^2 - 1,$$

$$9^2 + 1 \leq N-250 \leq 10^2,$$

or $332 \leq N \leq 350$.

B-2 (28, 8, 49, 0, 0, 0, 0, 56, 10, 39, 16)

Solution. Answer: For the real numbers c with $c < 243/8$.

The constant term and the coefficient of x in a quartic $p(x)$ are irrelevant in determining whether there is a line intersection $y = p(x)$ in four points. We may also replace $p(x)$ by $p(x - \alpha)$ for any real α . Thus, we may replace the given quartic $p(x) = x^4 + 9x^3 + cx^2 + 9x + 4$ with $p(x - 9/4) = x^4 + (c - 243/8)x^2 + \dots$, and drop the last two coefficients (we need never calculate them).

The problem then is to determine the values of c for which there is a straight line that intersects $y = x^4 + (c - 243/8)x^2$ in four distinct points. The result is now apparent from the shapes of the curves $y = x^4 + ax^2$. For example, we may note that when $a < 0$, this “W-shaped” curve has a relative maximum at $x = 0$, so that horizontal lines $y = -\varepsilon$ for small positive ε intersect the curve in four points, while for $a \geq 0$, the curve is always concave upward, so that no line can intersect it in more than two points.

B-3 (27, 10, 8, 5, 0, 0, 0, 2, 45, 49, 15, 45)

Solution. The desired set is $(-\infty, 1)$.

To show this, first note that if $k > 1$ were in the set, then $k = 1$ would also be in the set. However, if f is any function of the form $f(x) = g(x)e^x$, where g is a positive, increasing, differentiable function bounded by 0 and 1 (for example, $g(x) = (1/\pi) \arctan x + \frac{1}{2}$), we have $f'(x) = e^x(g'(x) + g(x)) > f(x)$ and $f(x) < e^x$ for all x , so $k = 1$ is not in the set.

On the other hand, if $f'(x) > f(x)$ for all x , then (since f is positive) we have

$$\frac{f'(x)}{f(x)} > 1 \quad \text{for all } x,$$

$$\int_0^x \frac{f'(t)}{f(t)} dt > \int_0^x 1 dt \quad \text{for all } x \geq 0,$$

$$\log(f(x)) > x + \log(f(0)) \quad \text{for all } x \geq 0,$$

$$f(x) > f(0)e^x \quad \text{for all } x \geq 0.$$

If k is any number less than 1, then for large enough x we will have $f(0)e^x > e^{kx}$ (since $f(0)$ is positive), which shows that k is in the set.

B-4 (15, 1, 4, 0, 0, 0, 0, 0, 5, 22, 71, 88)

Solution 1. From experimentation (and then an easy induction on n) we see that A^n has the form

$$A^n = \begin{pmatrix} a_n & b_n \\ 2b_n & a_n \end{pmatrix}$$

with a_n odd, and, since $\det A^n = 1$, we have $a_n^2 - 1 = 2b_n^2$. Thus $a_n - 1$ divides $2b_n^2$, so that $d_n = \gcd(a_n - 1, b_n) \geq \sqrt{(a_n - 1)/2}$. Since $\lim_{n \rightarrow \infty} a_n = \infty$ (e.g., $a_n > 3a_{n-1}$), the result follows.

Solution 2. Define the sequence r_0, r_1, r_2, \dots by $r_0 = 0$, $r_1 = 1$, and $r_k = 6r_{k-1} - r_{k-2}$ for $k > 1$. We first show by induction on k that

$$A^n = I = r_{k+1}(A^{n-k} - A^k) - r_k(A^{n-k-1} - A^{k+1}) \quad \text{for } k \geq 0. \quad (1)$$

This is clear for $k = 0$ and, for the inductive step, using $A^2 - 6A + I = 0$ (the characteristic equation), we have

$$\begin{aligned} & r_{k+1}(A^{n-k} - A^k) - r_k(A^{n-k-1} - A^{k+1}) \\ &= r_{k+1}((6A^{n-k-1} - A^{n-k-2}) - (6A^{k+1} - A^{k+2})) - r_k(A^{n-k-1} - A^{k+1}) \\ &= (6r_{k+1} - r_k)(A^{n-k-1} - A^{k+1}) - r_{k+1}(A^{n-k-2} - A^{k+2}) \\ &= r_{k+2}(A^{n-k-1} - A^{k+1}) - r_{k+1}(A^{n-k-2} - A^{k+2}). \end{aligned}$$

Applying (1) with $k = \lfloor n/2 \rfloor$, we obtain

$$A^n - I = \begin{cases} r_{n/2}(A^{n/2+1} - A^{n/2-1}), & \text{if } n \text{ is even,} \\ (r_{(n+1)/2} + r_{(n-1)/2})(A^{(n+1)/2} - A^{(n-1)/2}), & \text{if } n \text{ is odd.} \end{cases}$$

In either case, the entries of $A^n - I$ have a common factor that $\rightarrow \infty$ since $\lim_{n \rightarrow \infty} r_n = \infty$ (e.g., $r_n > 5r_{n-1}$ for $n > 1$).

Solution 3. We know that the entries of A^n are each of the form $\alpha_1 \lambda_1^n + \alpha_2 \lambda_2^n$ where $\lambda_1 = 3 + 2\sqrt{2}$ and $\lambda_2 = 3 - 2\sqrt{2}$ (the eigenvalues of A). So, using the entries for $n = 1, 2$, we derive

$$A^n = \begin{pmatrix} \frac{\lambda_1^n + \lambda_2^n}{2} & \frac{\lambda_1^n - \lambda_2^n}{2\sqrt{2}} \\ \frac{\lambda_1^n - \lambda_2^n}{\sqrt{2}} & \frac{\lambda_1^n + \lambda_2^n}{2} \end{pmatrix}.$$

Observing that $\lambda_i = \mu_i^2$, where $\mu_1 = 1 + \sqrt{2}$ and $\mu_2 = 1 - \sqrt{2}$, we see

$$\begin{aligned} d_n &= \gcd\left(\frac{\lambda_1^n + \lambda_2^n}{2} - 1, \frac{\lambda_1^n - \lambda_2^n}{2\sqrt{2}}\right) \\ &= \gcd\left(\frac{(\mu_1^n - \mu_2^n)^2}{2}, \frac{(\mu_1^n - \mu_2^n)(\mu_1^n + \mu_2^n)}{2\sqrt{2}}\right) \\ &= \left(\frac{\mu_1^n - \mu_2^n}{\sqrt{2}}\right) \gcd\left(\frac{\mu_1^n - \mu_2^n}{\sqrt{2}}, \frac{\mu_1^n + \mu_2^n}{2}\right) \end{aligned}$$

since $(\mu_1^n - \mu_2^n)/\sqrt{2}$, and $(\mu_1^n + \mu_2^n)/2$ are rational integers. As $|\mu_1| > 1$ and $|\mu_2| < 1$, we conclude $\lim_{n \rightarrow \infty} (\mu_1^n - \mu_2^n) = \infty$. Hence, $\lim_{n \rightarrow \infty} d_n = \infty$.

Comment: The proof extends to establishing the same result for integral matrices $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ of determinant 1 and $|\text{trace}(A)| > 1$ (the latter to guarantee $r_n \rightarrow \infty$ where $r_n = \text{trace}(A)r_{n-1} - r_{n-2}$). A similar argument gives the same conclusion for the entries of $A^n - I$.

B-5 (11, 4, 4, 0, 0, 0, 0, 32, 10, 15, 130)

Solution. For any $\alpha \geq 0$ and any positive integer k , we have

$$f_\alpha^k(n^2) = \lfloor \alpha \lfloor \alpha \lfloor \cdots \lfloor \alpha n^2 \rfloor \cdots \rfloor \rfloor \leq \lfloor \alpha \cdot \alpha \cdots \alpha n^2 \rfloor = f_{\alpha^k}(n^2),$$

so it is enough to show that there exists an $\alpha \geq 0$ such that for $1 \leq k \leq n$,

$$f_\alpha^k(n^2) \geq n^2 - k \quad \text{and} \quad \alpha^k n^2 < n^2 - k + 1.$$

For $k = 1$, the first of these two inequalities yields $\alpha n^2 \geq n^2 - 1$; we will show that $\alpha = (n^2 - 1)/n^2 = 1 - 1/n^2$ will do. Using this value of α , we use induction on k to show that $f_\alpha^k(n^2) \geq n^2 - k$ for $1 \leq k \leq n$; in fact, if this holds for k , we

have

$$\begin{aligned}
 f_{\alpha}^{k+1}(n^2) &\geq f_{\alpha}(n^2 - k) \\
 &= \left\lfloor \left(1 - \frac{1}{n^2}\right)(n^2 - k) \right\rfloor \\
 &\geq \left\lfloor \left(1 - \frac{1}{n^2 - k}\right)(n^2 - k) \right\rfloor \\
 &= n^2 - (k + 1)
 \end{aligned}$$

completing the induction.

To show that $\alpha^k n^2 < n^2 - k + 1$, note that this inequality is clear when $n = 1$ and hence $k = 1$, $\alpha = 0$; for $n > 1$, the inequality is equivalent to

$$\begin{aligned}
 \alpha^k &< 1 - \frac{k-1}{n^2} \\
 \left(\frac{n^2-1}{n^2}\right)^k &< 1 - \frac{k-1}{n^2}, \\
 \left(\frac{n^2}{n^2-1}\right)^k &> \frac{1}{1 - \frac{k-1}{n^2}}, \\
 \left(\frac{n^2}{n^2-1}\right)^{k-1} &< \frac{n^2-1}{n^2} \cdot \frac{1}{1 - \frac{k-1}{n^2}} = \frac{n^2-1}{n^2-k+1}.
 \end{aligned}$$

Now,

$$\left(\frac{n^2}{n^2-1}\right)^{k-1} = \left(1 + \frac{1}{n^2-1}\right)^{k-1} \geq 1 + \frac{k-1}{n^2-1} = \frac{n^2+k-2}{n^2-1},$$

and it is easy to see by cross-multiplication that for $1 \leq k \leq n$,

$$\frac{n^2+k-2}{n^2-1} > \frac{n^2-1}{n^2-k+1},$$

completing the proof.

B-6 (14, 11, 1, 0, 0, 0, 0, 16, 10, 50, 104)

Solution. Observe that $n_a \equiv a \pmod{100}$ and $n_a \equiv 2^a \pmod{101}$.

Suppose $n_a + n_b \equiv n_c + n_d \pmod{10100}$. Then $n_a + n_b \equiv n_c + n_d \pmod{101}$, so

$$2^a + 2^b \equiv 2^c + 2^d \pmod{101}. \tag{1}$$

Also, $n_a + n_b \equiv n_c + n_d \pmod{100}$, so $a + b \equiv c + d \pmod{100}$, and therefore, by Fermat's Theorem (since 101 is prime), $2^{a+b} \equiv 2^{c+d} \pmod{101}$. That is,

$$2^a \cdot 2^b \equiv 2^c \cdot 2^d \pmod{101}. \tag{2}$$

From (1) and (2), we see that $\{2^a, 2^b\}$ and $\{2^c, 2^d\}$ are the same set modulo 101,

namely, the set of roots of the quadratic polynomial $(x - 2^a)(x - 2^b) = x^2 - (2^a + 2^b)x + 2^a 2^b = (x - 2^c)(x - 2^d)$ in the field \mathbf{Z}_{101} . To see that $\{a, b\} = \{c, d\}$, it suffices to show that the numbers 2^a for $a \in \{0, 1, \dots, 99\}$ are distinct modulo 101. That is, we need to show that the order of 2 modulo 101 is precisely 100. For this, it suffices to show that $2^{20} \not\equiv 1 \pmod{101}$ and $2^{50} \not\equiv 1 \pmod{101}$. We have $2^{10} = 1024 \equiv 14 \pmod{101}$, so that $2^{20} \equiv 14^2 \equiv -6 \pmod{101}$, from which $2^{50} \equiv 2^{20} 2^{20} 2^{10} \equiv 36 \cdot 14 \equiv -1 \pmod{101}$.

Klosinski / Alexanderson:

Department of Mathematics

Santa Clara University

Santa Clara, CA 95053

Larson:


Department of Mathematics

St. Olaf College

Northfield, MN 55057

PICTURE PUZZLE

(from the collection of Paul Halmos)



Hint: He doesn't look the same now as he did in 1951.

(see page 690)

Mathematics: Questions and Answers

Beno Eckmann
*Translated by Peter Hilton**

The International Congress of Mathematicians, which takes place every 4 years and which is being held this time in Zürich, opens today. It brings together some 3000 mathematicians, active in research and university teaching, from all over the world. Not only in view of the (temporary) inundation of the city with this particular species of scientist but also at other times is the question often asked, what do mathematicians really do? In what follows I will try to give some small insight into the nature and processes of this science.

Fermat's Theorem. In June 1993 a sensational report went round the mathematical world; by electronic mail it reached even faraway universities, academies and colleges with lightning speed. The famous 350-year-old Fermat Theorem had been proved. To the surprise of most mathematicians this report was also published in many non-specialist media, thereby reaching a broad public. The *New York Times* devoted a front page article to it and Andrew Wiles (Princeton), who had announced the proof¹, along with those who had prepared the way, especially Gerhard Frey (Essen) and Kenneth Ribet (Berkeley), became as famous overnight as the stars of the arts and sport. The problem, unsolved for 350 years, seemed to exercise as strong a fascination for laymen as for specialists.

For once it was neither the powerful technico-scientific applications nor the attractive coloured computer pictures and graphics which excited a large public, but the actual mathematical process—and this with an unprecedented intensity. It was therefore only to be expected that, over the ensuing days, we mathematicians were bombarded with questions from all sides. Did we take the opportunity to make people, near and far, more familiar with our science? For the questions, on the whole, went to the heart of the matter.

The basic underlying question in Fermat's Theorem should be explained for the sake of completeness. The equation $x^2 + y^2 = z^2$ has many solutions in integers, for example, $x = 3$, $y = 4$, $z = 5$. On the other hand, the equation $x^3 + y^3 = z^3$ has no solutions in integers, and Fermat asserted, in 1635, that he could prove that the equation $x^n + y^n = z^n$, for $n > 2$, has no solutions in integers (except, of

* The original text of this article, in German, appeared in the Swiss newspaper *Neue Zürcher Zeitung* on August 3, 1994, to mark the opening of the International Congress of Mathematicians. It was suggested to the author, Beno Eckmann, that an English version would be welcomed and would reach a wider public. The translation was undertaken by Peter Hilton.

¹At the time of writing (8/3/94—translator) it appears to experts that there is a gap in the complicated chain of inferences constructed by A. Wiles. This does not imply false reasoning, but rather that the argument must be supplemented. The great achievement of Wiles is only marginally affected by this.

Added by translator (2/21/95)—It has now been announced by Andrew Wiles and R. L. Taylor (Cambridge), and verified by colleagues, that the gap has been filled.

course, $x = y = z = 0$). We may doubt whether he really had a proof. The problem seems simple and innocuous; it has aroused the interest of many amateur and professional mathematicians over the centuries, and many false proofs have been offered.

On the other hand the assertion has been proved for very many values of the exponent n ; up to last year, this included all values of n up to 4,000,000. It is noteworthy that the very profound methods which were developed to do this have had a decisive influence on modern mathematics; the theory of the so-called *algebraic numbers*, from which so many general ideas stem, arose primarily from these efforts. And now we know, provided Wiles' arguments are found to be watertight, that the assertion is valid for *all* n .

All this leaves the layman somewhat perplexed—so simple a statement and yet so difficult to prove? And do professional mathematicians occupy themselves with such things—and get paid for doing so? The further questions, below, give us the opportunity, in this respect, to correct some misapprehensions.

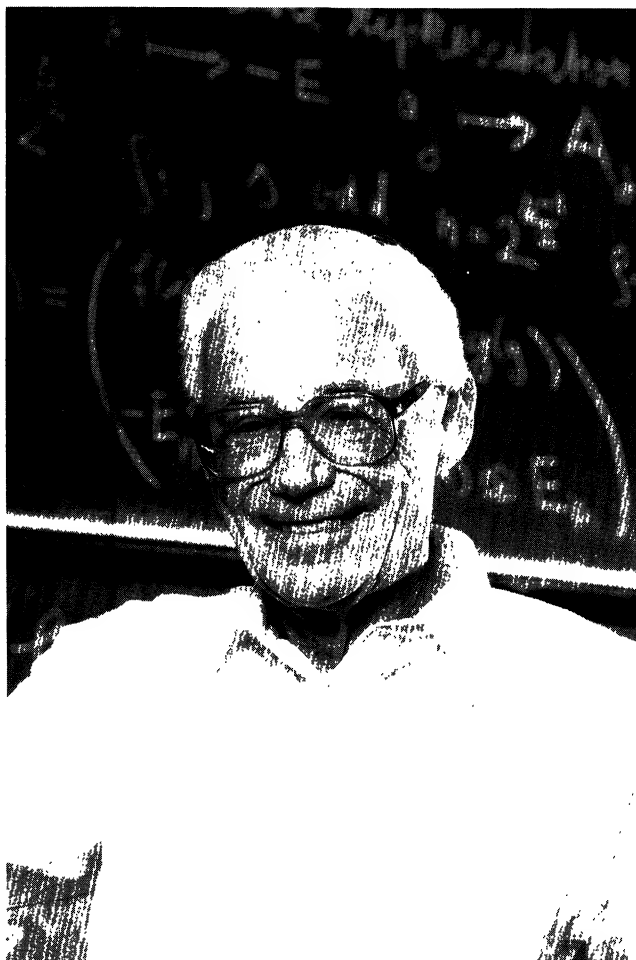
An invisible part of our culture. *What have we gained? What are the consequences for our world of the proved result?*

Here we must in all honesty reply: we have gained nothing. That the theorem has been proved has no consequences, even for number theory itself. But—does one pose such a question in the face of a masterwork of art or an impressive achievement in sport? Mathematics is, like the arts, a part of our *cultural tradition*, and has always, in ancient and modern times, obtained its justification from this fact. But, in contrast to the arts and sport, mathematics has no general public. Its assertions, as we must recognize, are immediately accessible only to a small circle; and the newer, the deeper, the more abstract the result, the narrower the circle will be. Thus mathematics can scarcely rely on making a resounding splash in the media, apart from very exceptional cases such as Fermat's Theorem.

But, of course, that is just one side of the story. On the other side stand the innumerable applications of mathematics which one meets everywhere. Mathematics has become an indispensable tool in the technico-scientific world of today, whether it is concerned with various kinds of calculation, with physics, chemistry, biology, medicine, meteorology, telecommunications etc. Even in simpler matters do all of us, knowingly or unknowingly, apply mathematical reasoning, when we speak of probability, extrapolation, analysis and interpretation of graphs, coding, averages and such like.

One does not, however, reflect that all the mathematical concepts, methods and results which are applied are *abstractions*, which had to be thought up. And even the solution of apparently 'frivolous' and useless problems à la Fermat—and many others originating in simple, practical questions—demand the elaboration of theoretical structures of great generality. The universal applicability of mathematics, which, as a rule, is neither intended nor foreseeable, seems to depend on those conceptions; a few examples to illustrate this will be cited below.

The relationship between these two very different aspects of mathematics is not easily comprehended. The instrument we employ for recognizing, describing, understanding and expressing by means of theoretical construction is mathematics, its language, its mode of thought, its results; that is, a structure of thought which is abstract and which is not primarily erected for this purpose. The applications bear witness to the power of mathematics, but are not its real motivation. The springs of



Beno Eckmann

“mathematization” seem to be of a very different kind. If we try to describe them, we need words like curiosity, thirst for knowledge, the impulse towards play.

A game then, pretentious and difficult, as all good games should be? In a certain sense, yes. But one knows that ultimately it has significance and effect and that places the motivation close to that of the artist. And, as in the arts, the criteria of value and rightness are not easily made precise. They include intensity, beauty and unity of the expression, the opening of new horizons, and insights which stem from a profound struggle to understand the problem. Even this remains inevitably restricted to the circle of the ‘initiated’. Thus is our art invisible to a wider public.

Mathematical proof. *Why prove something which is known to be correct in 4,000,000 cases, and more besides? Wouldn't one regard this, in any other endeavour, itself as “proof”?*

Here we must again go further back and, above all, insist that all those mathematical concepts, which are daily and hourly in action, find no place in the

real world we observe. The apparently simplest things like a straight line, 3-dimensional space, whole numbers, probability are creations of the human spirit, to say nothing of real or complex numbers, groups, vector spaces, integrals etc. Whether all these exist outside our thoughts or not, i.e., whether it is a matter of discovering or inventing, is also a bone of contention among mathematicians—but irrelevant here.

Certainly these ideas arise originally from our observations and experience, mainly in the domain of geometry and physics on the one hand and numbers and counting on the other. But first must come the complete abstraction, the release from reality, to form from that experience a *mathematical object*. This is only defined by its combinatorial properties, which vary from case to case and which satisfy certain axioms; essential here is the structure of *mutual relations*. In the framework thus established we apprehend, guided by intuition and experiment, relationships, results, theorems. Whether they are correct one can only determine by a strictly logical analysis of the proof—otherwise one does not know whether they are valid. Experience shows that intuition may lead us astray. So long as we have no proof of Fermat's Theorem, we cannot be sure that integer solutions do not exist for large values of the exponent n .

Concerning the multiplicity of applications of mathematical structures and results, this obviously stems from their *universality*, their independence from concrete objects. Whether it concerns the forecast of an eclipse of the sun or the moon, the mathematical design of a bridge, the formulation of cosmological theories, the schemata of the physics of elementary particles, or the analysis of computer tomograms, there are always abstract, mathematical tools behind it, far removed from any reality. It would be very dangerous to apply them if one were not sure of their validity.

No Nobel Prize. *Will Andrew Wiles receive the Nobel Prize?*

There is no Nobel Prize for mathematicians; this doesn't seem to be well-known, but it gives rise to speculation. Many explanations circulate, stories about conflicts between Nobel and a prominent mathematician of the time, and much more besides; as the President of the Nobel Committee once expressed it, not all these stories can be true. We don't know the reason, we can only conjecture: mathematics was simply *forgotten*. As so often happens, it was seen as a tool, which is simply to hand and which we apply; the mathematician's task is merely to carry out the necessary calculations. Even today when we generally recognize the significance of mathematics, people know very little of its true nature and inner beauty—because the research takes place within a narrower circle and is invisible from the outside. The non-mathematician sees only the tip of the iceberg. What is beneath? There lies this difficult and scarcely intelligible process of creating mathematical ideas and structures out of the vague experience and intuition of our environment, putting them to work and recognizing their connections; and even struggling with *totally unexpected consequences* of our own thinking. These are consequences which can give rise to far-reaching applications, from which further problems arise which call for new solutions or demand more new ideas.

An example which especially well illustrates how mathematical thought emerges from the depths to break surface is the discovery of electromagnetic waves, certainly one of the most important events in the history of science and modern

mankind. The credit should be given to the physicists James Clark Maxwell (1831–1879) and Heinrich Hertz (1857–1894); but it rests heavily on mathematical theories which had been developed much earlier for other reasons (analysis, the wave equation), and which showed that the Maxwell-Heaviside equations lead inevitably to *waves*— and this was experimentally verified by Hertz.

Similarly much else came in unexpected ways to be applied to the physical world: group theory, developed by Galois to study the solution of algebraic equations, has been applied to the elucidation of atomic spectra; Boolean algebra, which stems from mathematical logic, is applied to electric circuit theory; the Radon transform has been applied to computer tomography; category theory to the design of automata and formal languages; differential geometry, topology and algebra to the new theoretical physics. Always there were completely different reasons for creating and formulating the mathematical concepts—or perhaps no other reason but the inner beauty of the conceptual construction?

What about the computer? *Can one not simply leave the difficult considerations involved in the Fermat proof to the computer?*

This question is often asked, with some justification. For it is known not only to those involved, but also to the outsiders, that this is the era of the computer, which has immeasurably increased the possibilities for applying mathematical thought to our world. Moreover, not only applied, but also *pure* mathematicians, are using the computer in the most intensive way, to experiment, to verify conjectures, to render complicated geometric situations intelligible, and to push through difficult algebraic manipulations. But none of this replaces strict conceptual proof; on the contrary, it, in fact, depends on its logical foundations.

Now in an article which appeared last year in the *Scientific American* the “Death of Proof” was announced². The text was very well documented and contained quotations from well-known mathematicians. Classical proofs within a conceptual framework were to be replaced by visualization and verification, naturally on a computer; the Fermat proof by Wiles was characterized as a “splendid anachronism”. The article released a flood of indignant protests, even from mathematicians quoted in the article. All were agreed that the actual situation had been completely misunderstood. Semistrict arguments lead to semitruths which are correct only with a certain probability, or even false (and for whose uncertain validity huge amounts of computer time must be financed).

One could ignore this if a danger did not present itself whose consequences could be worse than one thinks. On the basis of such thinking a worldwide, fundamental restructuring of mathematics education could be proposed, which would replace everything by *interdisciplinary games* on the computer. It appears that already textbooks and software in this direction have been prepared, and here too certain reformers are following the same trend. Thus would the growing generations believe what they see on the screen, without knowing that “nothing has been proved”. And the experience of the inner beauty of mathematical thought would be withheld from them. Mathematics must be used according to its true nature, abstract, valid within a strict context, universal, and, precisely for that reason, eminently practical.

²John Horgan, The death of proof, *Scientific American*, October 1993.

So do the words of Hermann Weyl³, uttered 50 years ago, take on a new urgency:

“We do not claim for mathematics the prerogative of a Queen of Science; there are other fields which are of the same or even higher importance in education. But mathematics sets the standard of objective truth for all intellectual endeavours; science and technology bear witness to its practical usefulness. Besides language and music it is one of the primary manifestations of the free creative power of the human mind, and it is the universal organ for world-understanding through theoretical construction. Mathematics must therefore remain an essential element of the knowledge and abilities we have to teach, of the culture we have to transmit, to the next generation.”

Eckmann:
Mathematik
ETH-Zentrum
CH-8092 Zürich
SWITZERLAND
eckmann@math.ethz.ch

Hilton:
Department of Mathematical Sciences
SUNY at Binghamton
P.O. Box 6000
Binghamton, NY 13902-6000

³From the first page of the *Collected Works of Hermann Weyl*, edited by K. Chandrasekharan (Springer Verlag, 1968).

It is not knowledge, but the act of learning, not possession but the act of getting there, which grants the greatest enjoyment. When I have clarified and exhausted a subject, then I turn away from it, in order to go into darkness again; the never-satisfied man is so strange —if he has completed a structure, then it is not in order to dwell in it peacefully, but in order to begin another. I imagine the world conqueror must feel thus, who, after one kingdom is scarcely conquered, stretches out his arms for others.

—*Karl Friedrich Gauss (1777–1855)*

Letter to Bolyai, 1808.

Answer to Picture Puzzle

(p. 684)

Alex Rosenberg.

Teaching Math More Effectively, Through Calculational Proofs

David Gries and Fred B. Schneider

Lower-level college math courses usually avoid using formalism, in both definitions and proofs. Later, when students have mastered definitions and proofs written largely in English, they may be shown how informal reasoning could be formalized, but the impression is left that such formalization is not worth the effort. The design of proofs is also not taught. Students see proofs and may be asked to develop a few themselves, but there is little or no discussion of principles or strategies for designing proofs.

Few are happy with the results of these courses. Generally, students' reasoning abilities are poor, even after several math courses. Many students still fear math and notation, and the development of proofs remains a mystery to most. In short, students are not being equipped with the tools needed to employ mathematics in solving new problems.

We believe that this state of affairs can be improved. This article describes our approach.

THE INADEQUACY OF INFORMAL PROOFS. A proof of a theorem should provide evidence for belief in the validity of the theorem, where the evidence consists of facts (e.g. previously proved theorems) and an explanation of how they interact to convince. A good presentation of a proof should clearly indicate the facts and explain how they are combined. It should also make the proof appear so obvious that readers can see how it was developed, can explain it to others, and perhaps can prove other theorems in a similar fashion.

Now look at the proof in Table 1, which was taken from a math text and is typical of informal proofs. First, note that this proof does not state the facts on which it rests. (For example, it says, "If $y \notin A$, then, since $y \in A \cup B$ we must have $y \in B$ ", but there is no reference to the theorem that justifies this inference.)

TABLE 1. Conventional Proof of $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

We first show that $A \cup (B \cap C) \subseteq (A \cup B) \cap (A \cup C)$. If $x \in A \cup (B \cap C)$, then either $x \in A$ or $x \in B \cap C$. If $x \in A$, then certainly $x \in A \cup B$ and $x \in A \cup C$, so $x \in (A \cup B) \cap (A \cup C)$. On the other hand, if $x \in B \cap C$, then $x \in B$ and $x \in C$, so $x \in A \cup B$ and $x \in A \cup C$, so $x \in (A \cup B) \cap (A \cup C)$. Hence, $A \cup (B \cap C) \subseteq (A \cup B) \cap (A \cup C)$.

Conversely, if $y \in (A \cup B) \cap (A \cup C)$, then $y \in A \cup B$ and $y \in A \cup C$. We consider two cases: $y \in A$ and $y \notin A$. If $y \in A$, then $y \in A \cup (B \cap C)$, and this part is done. If $y \notin A$, then, since $y \in A \cup B$ we must have $y \in B$. Similarly, since $y \in A \cup C$ and $y \notin A$, we have $y \in C$. Thus, $y \in B \cap C$, and this implies $y \in A \cup (B \cap C)$. Hence $(A \cup B) \cap (A \cup C) \subseteq A \cup (B \cap C)$. The theorem follows.

Second, it is difficult to see precisely how the facts interact—the sequence and subsequences of inferences and all the case analyses in the proof cannot be easily digested. The structure of the proof is hidden by all the verbiage. One case analysis is presented in two paragraphs and others by sequential sentences within a paragraph; however, sequential sentences are also used to define steps common to all cases. Finally, this proof yields little insight into its development—how did it arise?

And yet, in spite of its inadequacies, this proof (and others like it) is held up as a model for students to emulate.

CALCULATIONAL PROOFS IN AN EQUATIONAL LOGIC. Our thesis is that mathematics and rigorous thinking can be taught more effectively by first teaching the design of rigorous proofs using a formal logic. However, the choice of logic and the accompanying method of proof is critical to success. In our experience, an equational logic, which is based on equality and Leibniz’s “substitution of equals for equals”, is most suitable because it has the following characteristics.

- Equational logic is easy to teach, since the style is already familiar to those who have had high-school algebra.
- Equational logic provides an alternative to reasoning in English. Rarely do proofs in equational logic parrot informal English arguments. Instead, proofs are *calculational*, in that they are developed by calculating using the rules of the logic, much as one calculates to solve a problem in high-school algebra. Further, principles and strategies can be used to help discover theorems and proofs.
- The rigorous use of equational logic need not lead to overwhelming complexity (as is the case with some logics). On the contrary, it is often a simplifying force. Typically, calculational proofs are shorter, simpler, and easier to remember than informal English proofs.
- Equational logic is versatile—it can be extended to a wide variety of mathematical domains.

Table 2 contains a calculational proof of theorem $p \vee q \equiv p \vee \neg q \equiv p$. Note that equivalence \equiv is being treated associatively, so that this theorem can be

TABLE 2. Equational proof of $p \vee q \equiv p \vee \neg q \equiv p$

	$p \vee q \equiv p \vee \neg q$
$=$	$\langle \text{Distr. of } \vee \text{ over } \equiv, p \vee (q \equiv r) \equiv p \vee q \equiv p \vee r \rangle$
	$p \vee (q \equiv \neg q)$
\equiv	$\langle \neg q \equiv q \equiv \text{false} \rangle$
	$p \vee \text{false}$
$=$	$\langle \text{Identity of } \vee, p \vee \text{false} \equiv p \rangle$
	p

viewed either as $(p \vee q \equiv p \vee \neg q) \equiv p$ or as $p \vee q \equiv (p \vee \neg q \equiv p)$. Symbol \equiv is used conjunctionally: $b = c = d$ is equivalent to $b = c \wedge c = d$.¹ Use of associativity of equivalence helps avoid formal detail without sacrificing rigor—our notation is designed with an eye to preventing complexity from overwhelming.

¹Operator \equiv is used for equality over booleans; $=$ is used for equality over any type, including boolean.

Each step of the proof in Table 2 has the following form.

$$\begin{aligned} & E[v := P] \\ = & \langle P = Q \rangle \\ & E[v := Q] \end{aligned}$$

Such a step shows equality of two formulas using the rule of “substitution of equals for equals”. The hint between the two formulas shows the equality being used in the substitution ($E[v := P]$ denotes expression E with every free occurrence of variable v replaced by expression P). Transitivity of equality allows us to conclude that the first and last formula of the proof of Table 2 are equal.

Notice that the proof format makes it easy to find the facts on which the proof depends—they are given within the angle brackets \langle and \rangle . Here, we have written out the full text of each fact, but we usually use the name or number of an already proved theorem.²

Explicit principles and strategies drove the development of the proof in Table 2. For example, one strategy for proving $P \equiv Q$ is to transform the more complicated of P and Q into the simpler one. In the proof, we viewed the formula to be proved as $(p \vee q \equiv p \vee \neg q) \equiv p$ and started with the more complicated, left-hand term. Second, the proof in Table 2 is “opportunity driven” or “forced”, in that at each step, the shape of the formula almost dictates in a unique way what substitution to make. Here, the shape of the first line of the proof cries out for simplification using distribution of \vee over \equiv . The second step is an equally obvious simplification, based on the shape of the formula.

Table 3 gives our calculational proof of distributivity of set union over set intersection. In contrast to the proof of Table 1, this proof exhibits all the good

TABLE 3. Calculational Proof of $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

Below, we prove $v \in A \cup (B \cap C) \equiv v \in (A \cup B) \cap (A \cup C)$. By Extensionality (the definition of equality of sets), we then conclude $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.

$$\begin{aligned} & v \in A \cup (B \cap C) \\ = & \langle \text{Definition of } \cup \rangle \\ & v \in A \vee v \in B \cap C \\ = & \langle \text{Definition of } \cap \rangle \\ & v \in A \vee (v \in B \wedge v \in C) \\ = & \langle \text{Distr. of } \vee \text{ over } \wedge \rangle \\ & (v \in A \vee v \in B) \wedge (v \in A \vee v \in C) \\ = & \langle \text{Definition of } \cup, \text{ twice} \rangle \\ & (v \in A \cup B) \wedge (v \in A \cup C) \\ = & \langle \text{Definition of } \cap \rangle \\ & v \in (A \cup B) \cap (A \cup C) \end{aligned}$$

qualities mentioned earlier. It refers to all the facts it uses (e.g. the definition of \cup). Its structure is simple, with each step being clearly delineated. And, it is based

²Formally, our logic consists of 15 axioms and 4 inference rules, and a theorem is either an axiom or a formula that is derived using the inference rules. The inference rules are Substitution of equals for equals (Leibniz), Transitivity of equality, Substitution, and Equanimity:

$$\begin{aligned} \text{Leibniz:} & \vdash P = Q \text{ then } \vdash E[z := P] = E[z := Q] \\ \text{Transitivity:} & \vdash P = Q, Q = R \text{ then } \vdash P = R \\ \text{Substitution:} & \vdash P \text{ then } \vdash P[z := Q] \\ \text{Equanimity:} & \vdash P, P \equiv Q \text{ then } \vdash Q \end{aligned}$$

on a strategy—one that is used over and over in mathematics: To prove something about operators (here, \cup and \cap), eliminate them using their definitions, perform some manipulation, and reintroduce the operators.

Anyone experienced in such calculational proofs will find the proofs of Tables 2 and 3 obvious and straightforward and will have no difficulty reproducing them. And, although these proofs are rigorous (and could be checked by a mechanical proof checker), complexity does not overwhelm.

Equational logic and the calculational approach can be extended to all domains typically taught in a first discrete math course—e.g. set theory, mathematical induction, a theory of integers, functions and relations, combinatorics, and recurrence relations. This is done by first defining the pure predicate calculus and then extending it by adding new types, presenting axioms that define the manipulative properties of the operations on those types, and building up a library of theorems.

A key to making rigor and formalism palatable is to keep notation consistent and uniform. Mathematics employs a number of different notations for quantification—see, for example, the left column of Table 4. We replace these different forms by a single notation for all quantifications. For any operator \star that is associative, is symmetric, and has an identity, the notation³

$$(\star i | R.i : P.i)$$

denotes the “accumulation” using operator \star of the values of expression $P.i$ over

TABLE 4. A Uniform Notation for Quantification

Conventional notation	Uniform notation
$\sum_{i=1}^3 i^2$	$(+i 1 \leq i \leq 3 : i^2)$
$(\forall i). 1 \leq i \leq 3 \Rightarrow b[i] = 0$	$(\wedge i 1 \leq i \leq 3 : b[i] = 0)$
$(\exists i). 1 \leq i \leq 3 \wedge b[i] = 0$	$(\vee i 1 \leq i \leq 3 : b[i] = 0)$
$\bigcup_{i=1}^3 S_i$	$(\cup i 1 \leq i \leq 3 : S_i)$

all values of i that satisfy range-predicate $R.i$. For example, Table 4 gives the conventional notation and a more uniform notation for four different quantifications. Other operators that can be used for \star are multiplication of integers, reals, and complex numbers, $b \cdot c$; union of sets, $S \cup T$; intersection of sets, $S \cap T$; minimum of two values, $b \downarrow c$ (if \downarrow does not have an identity, axioms and theorems that deal with a *false* range $R.i$ are not applicable); maximum of two values, $b \uparrow c$; and greatest common divisor, $b \text{ gcd } c$.

With a single notation, scope, free occurrence of a variable, and bound occurrence of a variable can be defined for all quantifications just once. More importantly, general axioms and theorems for manipulating quantifications can be introduced. The issue of quantification is thus simplified.

After introducing rules for quantification, it is easy to introduce pure predicate calculus. Operators \wedge and \vee are associative, are symmetric, and have identities,

³Bound variable i can be annotated with a type to indicate the range of values it may assume. A discussion of types is outside the scope of this article. Also, we write $R.i$ to denote application of function R to argument i ; eliminating the traditional parentheses avoids clutter.

so $(\wedge i|R.i : P.i)$ and $(\vee i|R.i : P.i)$ make sense. The first is universal quantification, more conventionally written as $(\forall i|R.i : P.i)$; the second is existential quantification, $(\exists i|R.i : P.i)$.

TEACHING THE CALCULATIONAL APPROACH. Equational propositional logic, along with preliminaries (e.g. the definition of textual substitution) can be taught to college freshmen in four weeks. During that time, students see many proofs and develop may themselves, in the calculational style. They also learn strategies and principles for designing proofs. As students develop a skill in proving theorems, they learn that attention to rigor may be simplifying force—and not an onerous burden.

Four weeks may seem like a long time to spend on propositional logic, but learning the calculational approach and gaining confidence in formal manipulation requires it and is worth it. Initially, most students are troubled by the prospect of uninterpreted manipulation. They want to think about the meanings of mathematical statements. Having meanings for objects is a “safety net”, which, students feel, prevents them from performing nonsensical manipulations. Unfortunately, the use of the “meaning” safety net does not scale well to complicated problems. Skill in performing uninterpreted syntactic manipulation does.

Students also have to be convinced that using formalism can be helpful. They must see first hand that a rigorous approach can help them solve problems they could not easily solve without it. This is possible with our approach. After just three days of learning equational logic, one can begin to attack the kinds of word problems that are found in Smullyan’s books, for example.

Once logic and proof have been thoroughly presented, other topics can be discussed—set theory, a theory of integers, and mathematical induction. Each topic is presented using the same calculational approach. In this manner, the notions of proof and proof style become the unifying force, the glue that binds together arguments in all domains. Discussion of informal versus formal presentations of proofs imparts deeper understanding of both, enabling students to deal more easily with math that they will see in later courses. For example, proof by contradiction in any domain is easily seen to be based on the theorem $p \equiv \neg p \Rightarrow \text{false}$ of propositional logic.

As an example of the greater understanding that rigor and precision allow, suppose we have proved the metatheorem that a formula P is a theorem iff the formula $(\forall x| : P)$ is a theorem. Then, the different ways in which theorems are expressed in texts can be discussed, and the following three statements can be seen to be equivalent. In the first, it is assumed informally that a and b are integers—perhaps this is mentioned in the accompanying prose; in the second, the type is given informally; in the third, the type is made formally explicit.

$$a + b = b + a$$

$$a + b = b + a \quad (\text{for } a, b \text{ integers})$$

$$(\forall a, b : \mathbb{Z} | : a + b = b + a)$$

To make rigor and formalism palatable, every new notation must be explained and rules must be given for manipulating it. Fear of formalism comes from having to use a formalism without knowing rules for its use, and attention in a class to such basic detail overcomes this fear. For example, traditionally, students are not shown rules for manipulating summations like $\sum_{i=1}^3 i^2$; consequently, they have

trouble with mathematical induction, where problems require manipulation of such summations.

The following example shows how attention to rigor and formal detail provides a measure of clarity that is impossible to obtain otherwise. Consider proving $b^{m+n} = b^m \cdot b^n$, for n, m natural numbers, by mathematical induction. Without formalizing quantification and having rules for manipulating it, no amount of informal explanation will clarify for students the different roles of m and n in the proof. However, $b^{m+n} = b^m \cdot b^n$ is equivalent to $(\forall m, n | 0 \leq n \wedge 0 \leq m : b^{m+n} = b^m \cdot b^n)$, which can be rewritten (using an axiom of quantification and the ability to name a formula) as

$$(\forall n | 0 \leq n : P.n) \quad \text{where } P.n : (\forall m | 0 \leq m : b^{m+n} = b^m \cdot b^n).$$

Now it is clear that n is the “induction variable” and that induction hypothesis $P.n$ is a universal quantification over m .

Further, once students understand quantification, they can prove the following—using a calculational proof. Let U be a set and $<$ a binary relation over U . Then $(U, <)$ admits induction iff $(U, <)$ is well founded. This theorem, which is rarely mentioned in informal presentations, gives deeper insight into induction.

When formal notations are presented properly, as a repository of the facts and a means of clarification, students begin to like formalism and to rely on it. It is the formalism that provides rules for judging between sound and unsound inference and that helps expose ambiguity and eliminate it.

DISCUSSION. The rigorous approach to teaching math has not, as yet, been accepted. Two criticisms are heard frequently: (1) students can’t handle rigor and formalism, and (2) teaching syntactic manipulation impedes understanding that a more semantic and informal approach provides.

Our own experience belies the first criticism; in fact, the criticism should go the other way. Teaching mathematics through informalism is like driving in a fog. One sees dim figures in the distance, and every once in a while some of them suddenly appear clearly, but usually everything is veiled and mysterious. It’s dangerous to drive in the fog, especially in a strange territory, and one must drive slowly. Even so, one may not always be sure where one is. Teaching rigor and precision, provided it is done without the veil of complexity interfering, burns away the fog, leaving everything crisp and clear and making it possible to drive faster and to enter uncharted lands.

We can rebut the criticism concerning semantics versus syntactics as well. An informal proof, like that in Table 1, can be translated into a proof in a natural-deduction or Hilbert-style logic. The resulting proof is every bit as syntactic as ours. The English proof is simply an informal version of a syntactic proof—and, as we have seen, a poor one at that. Therefore, the informal proof has no more meaning or semantics than a formal calculational proof.

Perhaps this criticism concerning semantics comes about because formal statements are sometimes difficult to understand. However, presenting a formal definition or theorem does not preclude giving alternative views as well. For example, a presentation of the axiomatic definition of set union can be supplemented with a Venn diagram, an English description, and an informal notion of evaluation. Nevertheless, it should be realized that for purposes of reasoning—constructing proofs—it is the axiomatic definition that is important. In fact, the axiomatic definition should be viewed as encoding all the meaning of the object being defined.

We also hear complaints that our approach suppresses intuition, that everything begins to appear mechanical. By “intuition” one usually means direct perception of truth or fact, independent of any reasoning process; keen and quick direct insight; or pure, untaught, noninferential knowledge (*Webster’s Encyclopedic Unabridged Dictionary*, 1989). There is simply no hope of teaching this—how can one teach something that is untaught, noninferential, and independent of any reasoning process? Of course, one can hope that students will develop an ability to intuit by watching instructors in math courses over the years. But this hit-or-miss prospect cannot be called *teaching* intuition.

On the other hand, a good part of mathematics is concerned with the opposite of intuition: with new and different reasoning processes that complement our ability to reason in English. This part of mathematics can be taught, and our approach to logic is an excellent vehicle for that task. Further, using the calculational approach to proofs, we are able to teach aids to discovery. In particular, with our disciplined, syntactic, proof style, we can teach principles and strategies whose application can indeed lead to the discovery of some (but not all) theorems and proofs. We have yet to see comparable principles and strategies for conventional English proofs.

Note that we are not against intuition; we have only stated that it cannot be taught. Moreover, we believe that discussing aids to discovery, as explained in the last paragraph, does not suppress intuition but goes further in aiding it than does the conventional method of teaching proofs.

New ideas in teaching are slow to catch on. People don’t like changing their habits—especially if it requires them to change their own way of thinking. However, current teaching methods are not exciting students or even educating them well, and alternatives should be seriously considered. Our approach bears looking into by all who want to teach mathematics effectively.⁴

Department of Computer Science
Cornell University
Upson Hall
Ithaca, NY 14853
gries@cs.cornell.edu
fbs@cs.cornell.edu

⁴The authors’ 500-page text *A Logical Approach to Discrete Math* (Springer Verlag, NY, 1993) uses the approach described in this article in teaching the usual topics in discrete math—logic, set theory, a theory of integers, induction, functions and relations, combinatorics, solving recurrence relations, and graph theory. The 300-page Instructor’s Manual contains other essays that concern the approach, as well as answers to the exercises. Together, the text and Instructor’s Manual contain over 700 calculational proofs, most of which are short and simple. Contact Gries at gries@cs.cornell.edu to obtain the Instructor’s Manual.

Some Problems Concerning Recurrence Sequences

G. Myerson and A. J. van der Poorten

There are questions about recurrence sequences that seem to crop up again and again. Plainly, though their answers are well known they are not known well. We endeavour to explain these answers in context so that they may become more widely known. The sequence $0, 1, -1, 2, -2, \dots$, in which each integer occurs exactly once, is a *recurrence sequence*; that is, it satisfies a linear, homogeneous recurrence relation with constant coefficients, namely,

$$a_n = -a_{n-1} + a_{n-2} + a_{n-3}.$$

It is not hard to produce a recurrence sequence in which each integer occurs exactly twice, or for that matter exactly n times, for any given n —we will show how to do this later. Can there be a recurrence sequence in which each integer occurs infinitely often? In which every rational number occurs? Every Gaussian integer? We will present the theory that enables us to answer these and many other questions about the range of a recurrence. At the pinnacle of this theory is the beautiful Skolem-Mahler-Lech Theorem, which deserves to be more widely known.

Let us first make some very general remarks about recurrence sequences. Suppose that the sequence a_0, a_1, \dots satisfies the relation

$$a_{h+n} = s_1 a_{h+n-1} + \dots + s_n a_h$$

for some complex numbers s_1, \dots, s_n and for $h = 0, 1, \dots$. Taking $h = 0$, we see that a_n is in the ring $\mathbb{Z}[a_0, \dots, a_{n-1}, s_1, \dots, s_n]$. An easy induction argument shows that, in fact, all the terms in the sequence belong to this ring. Thus, the entire sequence belongs to a ring finitely generated over \mathbb{Z} , the integers.

It follows immediately that it is impossible for every rational number to occur in a recurrence sequence, as the rationals are not contained in any finitely generated extension of the integers.

A little more is true. If we are dealing with rational (or even algebraic) numbers then it makes sense to speak of a common denominator d_0 for the numbers a_0, \dots, a_{n-1} and a common denominator d for s_1, \dots, s_n . It is clear by induction (or immediate by what we say below) that then the numbers $d_0 d^h a_h$ all are integers.

1. THE SKOLEM-MAHLER-LECH THEOREM. To settle the other questions raised in our opening paragraph, we must invoke the theorem of Skolem, Mahler, and Lech;

Theorem A. *If a_0, a_1, \dots is a recurrence sequence, then the set of all k such that $a_k = 0$ is the union of a finite (possibly empty) set and a finite number (possibly zero) of full arithmetic progressions.*

Here, a *full arithmetic progression* means a set of the form $\{r, r + d, r + 2d, \dots\}$ with $0 \leq r < d$. To illustrate, consider the sequence given by the recurrence $a_{n+6} = 6a_{n+4} - 12a_{n+2} + 8a_n$, with initial conditions $(a_0, \dots, a_5) = (8, 0, 9, 0, 8, 0)$;

$$8, 0, 9, 0, 8, 0, 4, 0, 0, 0, 16, 0, 128, 0, \dots$$

The set of k such that $a_k = 0$ is the union of the finite set $\{8\}$ and the full arithmetic progression $\{1, 3, 5, \dots\}$; in fact, the sequence is given by $a_n = 0$ if n is odd, $a_n = (n - 8)^2 2^{(n-6)/2}$ if n is even.

As so often happens, the proof of the theorem involves notions rather more sophisticated than its statement; so much so, that we can give only the barest sketch here. We will first tell the story of *generalized power sums* and make some introductory remarks about *p-adic analysis*, two of the important notions underlying the proof of the Skolem-Mahler-Lech Theorem, and of interest in their own right. The reader who is willing to accept the theorem on faith and eager to see the solutions of the problems posed above can read enough of the next section to understand the notation and then skip to Section 6 for the applications. The ambitious reader may then go on to the more advanced exposition written by the second author [vdP], or the detailed proof of Theorem A given by Cassels [Cas].

We note in passing that, for our purposes, $7, 0, 0, 0, \dots$ is not a recurrence sequence; the recurrence must hold from the start. The reader will experience no difficulty in extending the results given here to recurrences that only kick in after one or more terms of a sequence.

2. GENERALIZED POWER SUMS. A *generalized power sum* $a(h)$, $h = 0, 1, 2, \dots$ is an expression of the shape

$$a(h) = \sum_{i=1}^m A_i(h) \alpha_i^h, \quad h = 0, 1, 2, \dots \quad (1)$$

with *roots* α_i , $1 \leq i \leq m$, which are distinct non-zero quantities, and *coefficients* $A_i(h)$ which are polynomials respectively of degree $n_i - 1$, for positive integers n_i , $1 \leq i \leq m$. The generalized power sum $a(h)$ is said to have *order* $n = \sum_{i=1}^m n_i$.

Set

$$s(X) = \prod_{i=1}^m (1 - \alpha_i X)^{n_i} = 1 - s_1 X - \dots - s_n X^n. \quad (2)$$

Then the sequence (a_h) with $a_h = a(h)$, $h = 0, 1, 2, \dots$ satisfies the recurrence relation

$$a_{h+n} = s_1 a_{h+n-1} + \dots + s_n a_h, \quad h = 0, 1, 2, \dots \quad (3)$$

To see this let $E: f(h) \mapsto f(h + 1)$ be the shift operator. Its properties include:

- (i) $E^n(f(h)) = f(h + n)$,
- (ii) $E(f + g) = E(f) + E(g)$, and
- (iii) for all complex α and β ,

$$(E - \alpha)(E - \beta) = (E - \beta)(E - \alpha) = E^2 - (\alpha + \beta)E + \alpha\beta.$$

We have

$$(E - \alpha_i)(A_i(h) \alpha_i^h) = A_i(h + 1) \alpha_i^{h+1} - A_i(h) \alpha_i^{h+1} = (\Delta A_i(h)) \alpha_i^{h+1},$$

where $\Delta A_i(h) = A_i(h + 1) - A_i(h)$ is a polynomial of lower degree than that of A_i . By induction, $(E - \alpha_i)^{n_i}(A_i(h) \alpha_i^h)$ is identically zero. Let P be the operator

given by $P = \prod_{i=1}^m (E - \alpha_i)^{n_i}$. It follows that

$$P(a(h)) = P \sum_{j=1}^m A_j(h) \alpha_j^h = \sum_{j=1}^m P(A_j(h) \alpha_j^h) = 0.$$

But

$$\begin{aligned} P(a(h)) &= (E^n - s_1 E^{n-1} - \cdots - s_n) a(h) \\ &= a(h+n) - s_1 a(h+n-1) - \cdots - s_n a(h). \end{aligned}$$

Thus generalized power sums correspond to the sequences satisfying the recurrence relations (3). They also correspond to the Taylor coefficients of power series expansions of rational functions. Indeed, it follows from the above that there is a polynomial $r(x)$, of degree less than n , so that the power series

$$\sum_{h=0}^{\infty} a_h X^h = \frac{r(X)}{s(X)} \quad (4)$$

is a rational function; to see this multiply by $s(X)$ and note the recurrence relation (3).

Conversely, suppose we are given a rational function (4) as above, and suppose $\deg r < \deg s$. A partial fraction expansion, together with the well-known identity

$$(1 - Y)^{-j} = \sum_{h=0}^{\infty} \binom{h+j-1}{j-1} Y^h,$$

yields

$$\frac{r(X)}{s(X)} = \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{r_{ij}}{(1 - \alpha_i X)^j} = \sum_{h=0}^{\infty} \left(\sum_{i=1}^m \sum_{j=1}^{n_i} r_{ij} \binom{h+j-1}{j-1} \alpha_i^h \right) X^h.$$

The combinatorial symbols displayed are polynomials of degree $j-1$ in h , so the coefficients of X^h , $h = 0, 1, 2, \dots$ are indeed the values of a generalized power sum as described.

Accordingly, results on generalized power sums are equivalent to corresponding results on the Taylor coefficients of power series expansions of rational functions.

Later, we will need to deal with exponential polynomials

$$a(z) = \sum_{i=1}^m A_i(z) \exp(z \log \alpha_i), \quad (5)$$

the continuations to \mathbb{C} of generalized power sums. These are the solutions of linear differential equations with constant coefficients. To be precise, with $D = d/dz$, (5) is annihilated by the differential operator $\prod_{i=1}^m (D - \log \alpha_i)^{n_i}$. The order of the exponential polynomial (5) is n , as for the corresponding generalized power sum.

It is plain that an exponential polynomial vanishes identically if and only if all its coefficients vanish. We see this readily by induction on the order. Indeed, a one term exponential polynomial $A(z) \exp(z \log \alpha)$ obviously vanishes identically if and only if $A(z)$ vanishes identically. If (5) vanishes identically, then so does $(D - \log \alpha_1)a(z)$, which has order $n-1$. By the induction hypothesis all its polynomial coefficients vanish; that is for all i the polynomials $(D - \log \alpha_1 + \log \alpha_i)A_i(z)$ vanish identically. Then, with the exception of the constant coefficient

of A_1 , all the polynomials A_i must vanish identically. Our remark about a one term exponential polynomial guarantees that also that coefficient vanishes, and we are done.

3. AN APPLICATION TO RECURRENCE SEQUENCES. Let us use the equivalence of recurrence relations and rational functions to produce a recurrence sequence in which each integer occurs exactly k times. We write $c^{(k)}$ for the block c, c, \dots, c of length k . The sequence $0^{(k)}, 1^{(k)}, -1^{(k)}, 2^{(k)}, -2^{(k)}, \dots$ clearly contains each integer exactly k times. The corresponding power series is

$$f(x) = x^k + \dots + x^{2k-1} - x^{2k} - \dots - x^{3k-1} + 2x^{3k} + \dots,$$

which factors as

$$x^k(1 + x + \dots + x^{k-1} - x^k - \dots - x^{2k-1})(1 + 2x^{2k} - 3x^{4k} + \dots).$$

This is a rational function, since $1 + 2x^{2k} + 3x^{4k} + \dots = (1 - x^{2k})^{-2}$. Thus, the original sequence is a recurrence sequence. With a bit more algebra, we see

$$f(x) = \frac{x^k}{(1-x)(1+x^k)^2} = \frac{x^k}{1-x+2x^k-2x^{k+1}+x^{2k}-x^{2k+1}},$$

so the sequence satisfies the relation

$$a_{h+2k+1} = a_{h+2k} - 2a_{h+k+1} + 2a_{h+k} - a_{h+1} + a_h,$$

together with the initial conditions $a_0 = \dots = a_{k-1} = 0, a_k = \dots = a_{2k-1} = 1, a_{2k} = -1$.

4. AN INTRODUCTION TO p -ADIC ANALYSIS. The absolute value function defined on the integers has the following properties;

- (i) $|x| \geq 0$ for all x ,
- (ii) $|x| = 0$ if and only if $x = 0$,
- (iii) $|xy| = |x| \cdot |y|$ for all x and y , and
- (iv) $|x + y| \leq |x| + |y|$ for all x and y .

There are other functions that have the same properties. Given any non-zero integer n , and any prime number p , we can write $n = p^a m$ with a and m integers, $a \geq 0$, and p and m relatively prime. Moreover, this expression is unique. Define the function $|\cdot|_p$ by $|n|_p = p^{-a}$. Thus, for example, $|35|_7 = \frac{1}{7}$, $|36|_7 = 1$, and $|36|_3 = \frac{1}{9}$. If by convention we take $|0|_p = 0$ for all p , then it is not hard to see that all the properties of $|\cdot|$ listed above hold for $|\cdot|_p$, for each p . In fact, the last property holds in a stronger form, namely,

$$(iv') \quad |x + y|_p \leq \max(|x|_p, |y|_p).$$

We call $|\cdot|_p$ the *p -adic absolute value*. Thinking about convergence with respect to this absolute value leads to some peculiar-looking formulas. For example, for the geometric series with first term 6 and common ratio 7, the equation

$$6 + 42 + 294 + 2058 + \dots = -1$$

is a blunder in the usual run of things, but quite correct in the 7-adics.

The p -adic absolute value is easily continued to a function on the rational numbers, enjoying properties (i) through (iv'); any rational x can be written as $x = p^a r/s$ with a, r , and s integers, and r and s both relatively prime to p . Thus, $|\frac{35}{36}|_7 = \frac{1}{7}$, and $|\frac{35}{36}|_3 = 9$.

Any rational x has a unique decimal expansion $x = \sum_{j=m}^{\infty} a_j 10^{-j}$ with a_j in $\{0, 1, \dots, 9\}$, the series converging in the usual absolute value. So, too, for each p , any rational x has a unique p -adic expansion $x = \sum_{j=m}^{\infty} a_j p^j$ with a_j in $\{0, 1, \dots, p-1\}$, converging in the p -adic absolute value. For example, in the 7-adics we have

$$\begin{aligned} \frac{17}{98} &= 7^{-2} \cdot \frac{17}{2} = 7^{-2} \left(9 + \frac{3}{1-7} \right) = 7^{-2} (2 + 1 \cdot 7 + 3 + 3 \cdot 7 + 3 \cdot 7^2 + \dots) \\ &= 5 \cdot 7^{-2} + 4 \cdot 7^{-1} + 3 + 3 \cdot 7 + 3 \cdot 7^2 + \dots, \end{aligned}$$

where we have used the geometric series expansion $\frac{1}{1-7} = 1 + 7 + 7^2 + \dots$.

Now consider the sequence 1, 1.4, 1.41, 1.414, 1.4142, ... of decimal approximations to the square root of two. If m is less than n , then the m th and n th terms of this sequence differ by less than 10^{-m} , a quantity which goes to zero as m increases. Such a sequence is called a *Cauchy sequence* (with respect to the usual absolute value). You can't help feeling such a sequence ought to have a limit, but this one doesn't—if you confine yourself to the rationals [Euc]. In analysis, it is useful for Cauchy sequences to have limits, so we embed the rationals in the larger set called the *reals*. Every real number has a decimal expansion, and every Cauchy sequence converges—we say the reals are *complete*. The details of the completion process can be found in many introductory analysis texts, for example [Gle].

Now consider the sequence $7, 7 + 7^2, 7 + 7^2 + 7^4, 7 + 7^2 + 7^4 + 7^8, \dots$. In 7-adic absolute value, the difference between the m th and n th terms in this sequence is $|7^{2^m} + \dots + 7^{2^{n-1}}|_7 = 7^{-2^m}$, which goes to zero as m increases. That is to say, this is a Cauchy sequence—if you view it 7-adically. It ought, then, to have a limit. It is not a geometric series, so it cannot have a rational limit. By a process formally identical to the construction of the reals, we embed the rationals in a larger set we denote \mathbf{Q}_p , and call the *p -adic rationals*. Every p -adic rational has a p -adic expansion, and the p -adic rationals are complete.

Back to the reals. There are non-constant polynomials which have real coefficients but no real roots, for example, $x^2 + 1$. If we extend the reals to a field containing a root of $x^2 + 1$, we obtain the complex numbers. *Mirabile dictu*, every non-constant polynomial with complex coefficients has a complex root. We say that the complex numbers are *algebraically closed*. The absolute value function is continued to the complex numbers by $|a + bi| = (a^2 + b^2)^{1/2}$. *Mirabile squared*, the complex numbers are complete (with respect to this absolute value). The important functions of calculus (rational, exponential, trigonometric, ...) can be continued to functions of a complex variable, and many problems about real functions become easier to handle in this larger domain.

Back to the p -adic rationals. They are not algebraically closed. For example, if α in \mathbf{Q}_7 were a root of $x^2 - 7 = 0$, we would have $|\alpha|_7 = 7^{-1/2}$, but if α had the 7-adic expansion $\alpha = \sum_{j=m}^{\infty} a_j 7^j$, we would have $|\alpha|_7 = 7^{-m}$, with m an integer. We can embed \mathbf{Q}_p in an algebraically closed field $\overline{\mathbf{Q}}_p$, although the miracle of “add one number, get the rest free” does not occur here. We can extend $|\cdot|_p$ to $\overline{\mathbf{Q}}_p$, but $\overline{\mathbf{Q}}_p$ is not complete. We can complete $\overline{\mathbf{Q}}_p$ to a field \mathbf{C}_p , and this field is the p -adic analogue of the complex numbers; it is complete and algebraically closed. There is a rich theory of analytic functions on \mathbf{C}_p , mirroring that on the complex numbers.

This material can be found in less telegraphic form in [Kob].

What is really going on is this: The set of all Cauchy sequences forms a ring once we define the operations termwise; that the set is closed under the operations is a consequence of the rules (i)–(iv). One defines the field of reals (respectively

p -adic rationals, according to the particular valuation defining ‘Cauchy’) to be this ring with sequences ‘with the same limit’ identified. What that means is that we take the subset of null sequences, those converging to 0, and notice again by the rules (i)–(iv) that this set is a maximal ideal in the ring of Cauchy sequences. Then the quotient ring is a field.

The ‘miracle’ of \mathbf{R} and \mathbf{C} actually is a rather special. It turns out that if a field \mathbf{F} is algebraically closed and if \mathbf{L} is a subfield of finite codimension in \mathbf{F} (in English: if \mathbf{F} is a finite-dimensional vector space over some field \mathbf{L}) then necessarily $[\mathbf{F}:\mathbf{L}] = 2$ (compare $[\mathbf{C}:\mathbf{R}] = 2$) and \mathbf{L} is an *ordered field*. That means that \mathbf{L} is the disjoint union of three sets N , $\{0\}$ and P with P closed under addition and multiplication and $N = -P$; P is of course the set of *positive* elements of \mathbf{L} . It turns out that \mathbf{L} can be ordered if and only if -1 is not a sum of squares. A complete orderable field is known as a *real field* and always is a subfield of codimension 2 in an algebraically closed field; for all this see for example [L], Chapter XI. By contrast, it is not hard to see that \mathbf{Q}_p is not an ordered field.

5. ON PROVING THE SKOLEM-MAHLER-LECH THEOREM. Recall that the terms of a recurrence are given by a generalized power sum,

$$a(h) = \sum_{i=1}^m A_i(h) \alpha_i^h, \quad h = 0, 1, 2, \dots \quad (1)$$

Given a positive number p , every h can be written uniquely as

$$h = r + (p-1)t, \text{ with } r = 0, 1, \dots, p-2 \text{ and } t = 0, 1, 2, \dots$$

If we write $a_{p,r}(t)$ for $a(h)$, we get

$$a_{p,r}(t) = \sum A_i(r + (p-1)t) \alpha_i^r \exp(t \log \alpha_i^{p-1}). \quad (6)$$

Now it can be shown that there exist primes p such that the logarithmic and exponential functions can be continued to analytic functions on \mathbf{C}_p —more accurately, on regions of \mathbf{C}_p large enough for the formula above to make p -adic sense for t in a closed set D containing the integers. Then $a_{p,r}(t)$ is a p -adic analytic function on D for $r = 0, 1, \dots, p-2$.

Suppose that there are infinitely many h such that $a(h) = 0$. Then there must be at least one r for which the analytic function $a_{p,r}(t)$ is zero for infinitely many integers t . Of course, there are *complex* analytic functions which are zero for infinitely many integer values of their argument; for example, $\sin \pi z$. This can occur because the integers are an unbounded set in \mathbf{C} . Things are different in \mathbf{C}_p , since the integers form a *bounded* set there; after all, $|n|_p \leq 1$ for all integers n . It turns out that a function (whether complex or p -adic) analytic on a closed, bounded region and with infinitely many zeros in that region must be identically zero. Thus, $a(r + (p-1)t)$ vanishes identically for all integer t , and in particular $a(h)$ is zero for all h in an arithmetic progression. This concludes our sketch of the proof.

It is a little strange that Theorem A should force us to enter the realm of p -adic analysis. Actually that can sort of (but not really) be avoided. It turns out that p must be selected so that $\alpha_i^p \equiv \alpha_i \pmod{p}$ for each i . Then (6) has no more than $n-1$ integer zeros (so certainly not infinitely many); otherwise it vanishes identically [RvdP]. The trouble is that there seems only to be a p -adic proof for the bound.

6. APPLYING THE SKOLEM-MAHLER-LECH THEOREM. So if there are infinitely many h such that $a(h) = 0$ then there must be at least one r for which the

analytic function

$$a_{p,r}(t) = \sum A_i(r + (p-1)t) \alpha_i^r \exp(t \log \alpha_i^{p-1})$$

vanishes identically. So, by our discussion at the end of §2, since the A_i are not identically zero, the $\log \alpha_i^{p-1}$ cannot all be distinct.

Indeed, the numbers α_i^{p-1} must coincide at least in pairs. Plainly $p-1$ is not arbitrary and depends only on the roots α_i .

Moreover, we see that the original function

$$a(z) = \sum_{i=1}^m A_i(z) \exp(z \log \alpha_i)$$

vanishes at all $z = r + t(p-1)$ with $t \in \mathbf{Z}$. As an aside we mention that then it follows that $a(z)$ must be the product of

$$\sin \frac{\pi}{p-1}(z-r) = \frac{1}{2i} \left(e^{\frac{\pi i}{p-1}(z-r)} - e^{\frac{-\pi i}{p-1}(z-r)} \right)$$

with some other exponential polynomial. In that sense a recurrence sequences has infinitely many zeros if and only if it is 'sinful'.

So, in particular (taking $l = p-1$, say) we have:

Proposition 1. *If a recurrence sequence vanishes infinitely often, then it vanishes on an arithmetic progression with a common difference 1 that depends only on the roots.*

Now suppose there is a number k such that $a(h) = k$ for infinitely many h . Let $b(h) = a(h) - k$. Then $b(h) = \sum_1^m A_i(h) \alpha_i^h - k \cdot 1^h$ is a generalized power sum with the same roots as $a(h)$ (and, possibly, the root 1 if it was not already a root of $a(h)$), hence the same l -values as $a(h)$, and $b(h)$ is zero whenever $a(h) = k$. Thus, $a(h)$ takes on the value k on an arithmetic progression with common difference l .

Now there are only l different complete arithmetic progressions of integers with common difference l . So we have established a principal remark of this note, namely,

Proposition 2. *The number of values that a recurrence sequence can take on infinitely often is bounded by some integer l that depends only on the roots.*

It follows immediately that there is no recurrence sequence in which each integer occurs infinitely often.

Nor is there a recurrence sequence in which every Gaussian integer occurs. For suppose a_h were such a sequence, and let $\sum_{h=0}^{\infty} a_h X^h = r(X)/s(X)$. Then

$$\sum_{h=0}^{\infty} \mathcal{R}(a_h) X^h = \mathcal{R} \frac{r(X)}{s(X)}.$$

Now it is easy to see that the real part of a rational function is again a rational function, so $\mathcal{R}(a_h)$ is a recurrence sequence, and it takes on every integer infinitely often. As we have seen, this cannot happen.

7. MULTIPLICITY: A GOOD QUESTION. We restrict ourselves to recurrence sequences of integers. By the results just explained an integer recurrence sequence either takes the value 0 infinitely many times, in which case it has special properties that allow us to say it is *degenerate*, or only finitely many times. Is there

a bound $\mu(n)$ so that a nondegenerate integer recurrence sequence of order n has at most $\mu(n)$ zeroes? Of course any given non-degenerate integer recurrence sequence has a bound on its number of zeroes. Our question is whether there is a *uniform* bound for the *multiplicity*, depending only on the order of the sequence.

It is obvious that $\mu(2) = 1$. (Truly. Give this a few minutes thought.) The bound $\mu(3) = 6$ is very much more difficult and has only been confirmed recently [Beu]. The extreme case is

$$a_{h+3} = 2a_{h+2} - 4a_{h+1} + 4a_h, \quad a_0 = a_1 = 0, a_2 = 1.$$

Its six zeroes are $a_0 = a_1 = a_4 = a_6 = a_{13} = a_{52} = 0$.

For larger n there are not even any worthwhile conjectures. The problem deserves some computer time, say at least so as to guess $\mu(4)$ (which is ≥ 9).

8. RECURRENCE. The question, whether there is a recurrence sequence in which each rational occurs, was raised in *Crux Mathematicorum* in October, 1989.

Proposition 2 was published in 1959 by Shapiro [Sha], and again some years later by Berstel and Mignotte [Ber]. The question, whether there is a recurrence sequence in which each Gaussian integer occurs infinitely often, was posed in *Crux Mathematicorum* in June, 1988, and repeated in October 1989. These sequences are recurrent in more ways than one! Indeed, Theorem A for recurrence sequences of algebraic numbers was first proved by Mahler in the 30's, based upon an idea of Skolem. Then, Lech published the result for general recurrence sequences in 1953. In 1956 Mahler published the same result, apparently independently (but later realized to his chagrin that he had actually reviewed Lech's paper some years earlier, but had forgotten it).

References not explicitly given here can be found in the survey [vdP].

REFERENCES

-
- [Ber] J. Berstel and M. Mignotte, *Deux propriétés décidables des suites récurrentes linéaires*, Bull. Soc. Math. France **104** (1976), 175–184, MR 54 #2576.
 - [Beu] F. Beukers, *The zero-multiplicity of ternary recurrences*, Compositio Math. **77** (1991), 165–177, MR 92a:11014.
 - [Cas] J. W. S. Cassels, *Local Fields*, Cambridge U. Pr., Cambridge, 1986, MR 87i:11172.
 - [Euc] Euclid, *The Thirteen Books of Euclid's Elements*, 2nd ed., T. L. Heath, ed., Dover, New York, 1956, MR 17-814.
 - [Gle] A. Gleason, *Fundamentals of Abstract Analysis*, Addison-Wesley, Reading, 1966, MR 34 #2378.
 - [Kob] N. Koblitz, *p-adic Numbers, p-adic Analysis, and Zeta-Functions*, Springer, New York, 1977, MR 57 #5964.
 - [L] Serge Lang, *Algebra*, Addison-Wesley, Reading, Mass., 1965, MR33 #5416.
 - [RvdP] A. J. van der Poorten and R. S. Rumely, *Zeros of p-adic exponential polynomials II*, J. London Math. Soc. **36** (1987), 1–15, MR 88m:11103.
 - [vdP] A. J. van der Poorten, *Some facts that should be better known; especially about rational functions*, in *Number Theory and Applications* ed. Richard A. Mollin, (NATO—Advanced Study Institute, Banff, 1988), Kluwer Academic Publishers, Dordrecht, 1989, pp. 497–528, MR 92k: 11011.
 - [Sha] H. N. Shapiro, *On a theorem concerning exponential polynomials*, Comm. Pure. Appl. Math. **12** (1959), 487–500, MR 22 #12078.

Centre for Number Theory Research
Macquarie University NSW 2109
Australia
gerry@mpce.mq.edu.au
alf@mpce.mq.edu.au

A Hyperbolic Plane Coloring and the Simple Group of Order 168

Dana Mackenzie

Monthly problem 10349, proposed by Raphael M. Robinson, reads:

The hyperbolic plane is tiled with equilateral triangles meeting seven at each vertex. Can the tiles be colored with seven colors in such a way that no two tiles of the same color meet, even at a vertex?

This paper will present the same solution of this problem in three different ways. The reason for this apparent surfeit of approaches, we shall see, is an unusual isomorphism between the two groups $PSL_2(\mathbb{Z}_7)$ and $SL_3(\mathbb{Z}_2)$. These groups are known to be isomorphic (see [2]), but I am not sure whether such a visual representation of this fact has been given before. We will use our coloring procedures to explain the isomorphism between $PSL_2(\mathbb{Z}_7)$ and $SL_3(\mathbb{Z}_2)$.

Solution 1. The simplest solution follows along the lines of the argument in [3, pp. 176–7] demonstrating the existence of the $(2,3,7)$ tessellation of the hyperbolic plane. Figure 1 illustrates a 7-coloring of the 3-holed torus T_3 such that no two

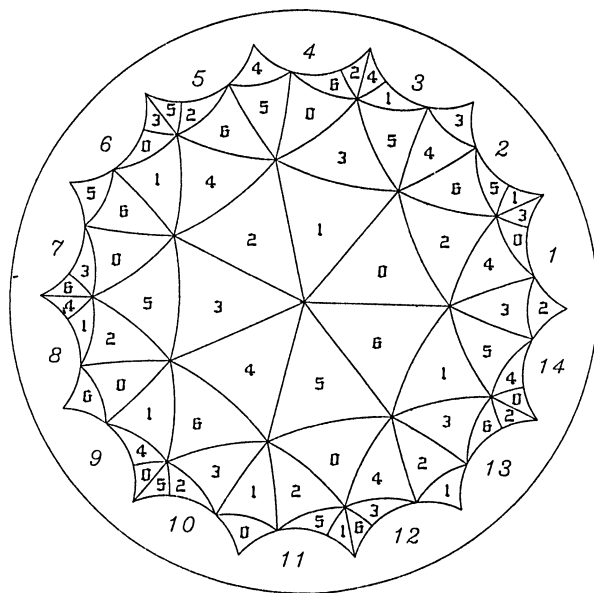


Figure 1. A seven-coloring of a hyperbolic polygon that can be “sewn together” to form a three-holed torus. Gothic numerals 0, 1, ..., 6 denote the colors of the triangular faces. Italic numerals 1, 2, ..., 14 indicate the ordering of the polygon, used to determine which sides are sewn together.

triangles of the same color share a vertex. The three-holed torus¹ is formed by identifying side $2i + 1$ with side $2i + 6 \pmod{14}$, if the sides are numbered as in Figure 1. (Note that some of the equilateral triangles are bisected by the lines which are “sewn together” by this procedure.) Lifting this coloring to the universal cover of T_3 , we obtain a 7-coloring of the hyperbolic plane with the desired property. ■

The trouble with Solution 1 is that, though it is concise, it gives no clue as to how the coloring in Figure 1 was generated. In addition, it relies on the somewhat serendipitous fact that the $(2, 3, 7)$ tiling can be generated by lifting a tiling of T_3 . In Solutions 2 and 3 we will explain how Figure 1 was derived, and we will assume only the standard facts that the $(2, 3, 7)$ tiling exists and that its (orientation-preserving) symmetry group is

$$\Gamma = \langle \alpha, \beta, \gamma \mid \alpha^7 = \beta^3 = \gamma^2 = \alpha\beta\gamma = e \rangle$$

The isometry α can be realized as the rotation by $-2\pi/7$ about the point P in Figure 2; β as the rotation by $-2\pi/3$ about the centroid of $\triangle PQR$; and γ as the rotation by π about the midpoint of segment PQ .

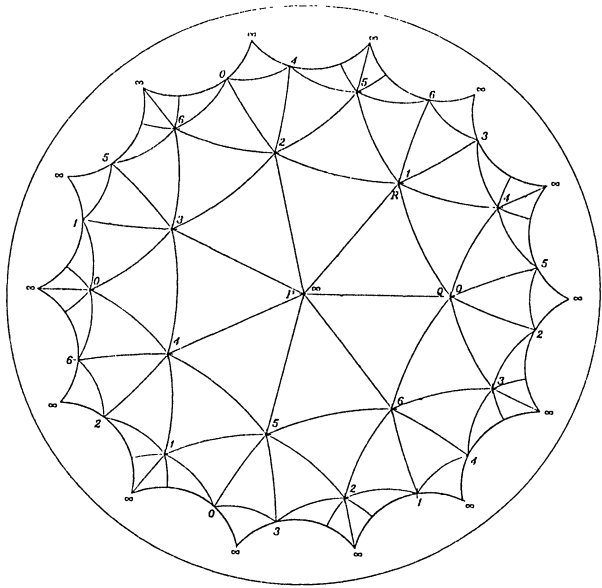


Figure 2. The tessellation of Figure 1, with the vertices labeled according to the method of Solution 2.

For any tessellation (in particular, the one under consideration), we define F to be the set of faces, E to be the set of edges, and V to be the set of vertices. We define a *coloring* of faces, a *highlighting of edges*, and a *labeling* of vertices, respectively, to be maps from F , E , and V to a finite set. In Solution 2, we will derive the coloring of Figure 1 from a labeling $L: V \rightarrow \mathbb{Z}_7 \cup \{\infty\}$ (where “ ∞ ” is an

¹This torus also called the *Klein quartic*, seems to be in vogue at the moment. In the past year it was described in the popular press as “famous . . . of almost mythological proportions” ([1]), and inspired the new sculpture “The Eightfold Way,” by H. Ferguson, unveiled in November 1993 at the Mathematical Science Research Institute in Berkeley, CA ([5]).

abstract symbol interpreted in much the same way as in complex analysis). In Solution 3 we will derive the same coloring from a highlighting $H: \mathbf{E} \rightarrow \mathbf{Z}_7$.

Solution 2. To begin, we define a homomorphism

$$\phi: \Gamma \rightarrow PSL_2(\mathbf{Z}_7)$$

by $\phi(\alpha) = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}$, $\phi(\beta) = \begin{bmatrix} 1 & -1 \\ 1 & 0 \end{bmatrix}$, $\phi(\gamma) = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$. (These matrices satisfy the same identities as α , β and γ , so we can be assured that such a homomorphism exists.) Recall that the matrices $\begin{bmatrix} a & b \\ c & d \end{bmatrix} \in PSL_2(\mathbf{Z}_7)$ act on the set $\mathbf{Z}_7 \cup \{\infty\}$ by linear fractional transformations:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}(x) \equiv \frac{ax + b}{cx + d} \pmod{7}.$$

Throughout this paper we will identify the matrices with the linear fractional transformations, for simplicity.

Now we can describe our labeling $L: \mathbf{V} \rightarrow \mathbf{Z}_7 \cup \{\infty\}$. Given $A \in \mathbf{V}$, let B be any adjacent vertex and let C be the unique adjacent vertex such that $\triangle ABC$ is in the tessellation and is positively oriented. Then there exists a unique isometry $\chi \in \Gamma$ such that $\chi(P) = A$, $\chi(Q) = B$, and $\chi(R) = C$. Then define $L(A) \equiv [\phi(\chi)](\infty)$.

A much simpler way to compute the labeling L is to apply the following recursive procedure. Begin by labeling the central vertex P “ ∞ ” (i.e. setting $L(P) = \infty$) and labeling the seven adjacent vertices “0” through “6,” beginning with Q and proceeding counterclockwise. Extend to a labeling $L(\chi)$ of all the vertices in the $(2, 3, 7)$ tessellation by applying the following rule:

Rule 1. *If the vertices of $\triangle ABC$ are already labeled, with $L(A) = a$, $L(B) = b$, $L(C) = c$, where a, b, c are distinct elements of $\mathbf{Z}_7 \cup \{\infty\}$, and if $\triangle ABC$ is an adjacent triangle, set $L(D) = d$, where d is the unique number in $\mathbf{Z}_7 \cup \{\infty\}$ such that $(a, b, c, d) = -1 \pmod{7}$.*

Here (a, b, c, d) denotes the cross ratio.

The uniqueness of d is a standard fact about cross ratios, but there is another reason that L may not be well-defined. Since each vertex belongs to seven different triangles, and each of these is adjacent to three different triangles, Rule 1 could give us many conflicting instructions about how to label a given vertex. The easiest way to show that it does not, in fact, lead to “conflicting instructions,” is to verify that our first, non-recursive procedure does give a labeling that satisfies Rule 1.

Lemma 1. *L is a well-defined labeling on \mathbf{V} and satisfies Rule 1.*

Proof: We need to show that L does not depend on the choice of the adjacent vertex B . Suppose we had chosen C (the next vertex, proceeding counterclockwise), thereby using the adjacent triangle $\triangle ACD$ instead of $\triangle ABC$. Then the isometry mapping $\triangle PQR \rightarrow \triangle ACD$ is $\chi' = \chi\alpha^{-1}$, so

$$L(A) = [\phi(\chi')](\infty) = [\phi(\chi\alpha^{-1})](\infty) = [\phi(\chi)][\phi(\alpha^{-1})](\infty) = [\phi(\chi)](\infty),$$

which agrees with the previous definition. Proceeding by induction counterclockwise about the vertex A , we conclude that $L(A)$ is independent of the choice of the adjacent vertex B .

To prove Rule 1 holds, we suppose that $\triangle ABC$ and $\triangle ABD$ are adjacent triangles in the tessellation. Without loss of generality, suppose that $\triangle ABC$ is located clockwise from $\triangle ABD$ in the cycle of seven triangles with vertex A . Let $\chi: \triangle PQR \rightarrow \triangle ACB$, then $\chi\beta^{-1}: \triangle PQR \rightarrow \triangle CBA$, $\chi\beta: \triangle PQR \rightarrow \triangle BAC$, and $\chi\alpha^{-1}\beta: \triangle DAB$. Hence

$$\begin{aligned} L(C) &= [\phi(\chi)][\phi(\beta^{-1})](\infty) = [\phi(\chi)](0), \\ L(B) &= [\phi(\chi)][\phi(\beta)](\infty) = [\phi(\chi)](1), \\ L(D) &= [\phi(\chi)][\phi(\alpha^{-1})][\phi(\beta)][\phi(\beta)](\infty) = [\phi(\chi)](2). \end{aligned}$$

Since $\phi(\chi)$ is a linear fractional transformation, it preserves cross ratios. Thus

$$(L(A), L(B), L(C), L(D)) = (\infty, 1, 0, 2) \equiv -1 \pmod{7}. \quad \blacksquare$$

Now we explain how to proceed from the vertex labeling L to a face coloring. Our coloring will be based on the fact that for any four distinct elements $a_1, \dots, a_4 \in \mathbf{Z}_7 \cup \{\infty\}$, the set of all values of the cross-ratio $(a_{\pi(1)}, a_{\pi(2)}, a_{\pi(3)}, a_{\pi(4)})$ as π ranges over all permutations is either $\{2, 4, 6\}$ or $\{3, 5\}$. In the latter case, the subgroup of permutations which fix the cross-ratio has index 2 and hence must be the alternating group A_4 . Thus, in particular, if $(a, b, c, d) \equiv 5 \pmod{7}$ then $(c, a, b, d) = (b, c, a, d) \equiv 5 \pmod{7}$ as well, while, for example, $(b, a, c, d) \equiv 3 \pmod{7}$.

We will define a triple (a, b, c) to be *positively oriented* if and only if there is a positively oriented triangle $\triangle ABC$ with labels $L(A) = a$, $L(B) = b$, and $L(C) = c$. Obviously this will hold if and only if there exists $f \in PSL_2(\mathbf{Z}_7)$ such that $f(a) = \infty$, $f(b) = 0$, and $f(c) = 1$. A linear fractional transformation which accomplishes this is

$$g = \begin{bmatrix} d - b & (b - d)c \\ d - c & (c - d)b \end{bmatrix}$$

This transformation can be represented by a matrix of determinant 1 if and only if $\det g$ is a quadratic residue. (If $\det g \equiv \lambda^2$, then $(1/\lambda)g$ is a matrix of determinant 1 which represents the same transformation.) Thus we may define a positively-oriented triple in a purely number-theoretic fashion: if $a, b, c \neq \infty$, then (a, b, c) is positively oriented if and only if $\det g = (a - b)(b - c)(c - a)$ is a quadratic residue $\pmod{7}$.

Similarly, one can check that (a, b, ∞) is positively oriented if and only if $(b - a)$ is a quadratic residue $\pmod{7}$.

Here is our rule for coloring the faces of the $(2, 3, 7)$ tessellation.

Rule 2. Define $C(\triangle ABC) = n$ if and only if

$$\{L(A), L(B), L(C)\} \subset S_n \quad \text{or} \quad \{L(A), L(B), L(C)\} \subset S_n^c,$$

where $S_n = \{n, n + 1, n + 3, \infty\}$. (As usual, all the additions are modulo 7. S_n^c denotes the complement of S_n in $\mathbf{Z}_7 \cup \{\infty\}$.)

An alternate definition for the sets S_n , and the reason for choosing these particular sets, is given in the next lemma.

Lemma 2. Let \mathcal{E} be the following collection of subsets of $\mathbf{Z}_7 \cup \{\infty\}$:

$$\mathcal{E} = \{\{a, b, c, d\} \mid (a, b, c, d) = 3 \text{ iff } (a, b, c) \text{ is positively oriented}\}.$$

Then $\mathcal{E} = \{S_n, S_n^c \mid 0 \leq n \leq 6\}$, and \mathcal{E} is a $(3, 4, 8)$ Steiner system.

(For the definition of a Steiner system, see, for example, [4].)

Proof: It is easily checked that the sets $\{0, 1, 3, \infty\}$ and $\{2, 4, 5, 6\}$ are in \mathcal{E} , and it follows for the remaining S_n 's, since they are obtained from these two by a translation (mod 7).

Next we note that for each of the $\binom{8}{3} = 56$ triples $\{a, b, c\} \subset \mathbf{Z}_7 \cup \{\infty\}$ there is a unique d which satisfies the condition in the definition of \mathcal{E} . This shows that \mathcal{E} is a Steiner system. Since each set $\{a, b, c, d\} \in \mathcal{E}$ can be generated from four different triples, there are 14 elements in \mathcal{E} . Thus the 14 sets of the form S_n and S_n^c must be all the elements of \mathcal{E} .

The fact that \mathcal{E} is a Steiner system is required for the coloring C defined in rule 2 even to make sense: we need to know that for each set of three vertex labels, there is only one and only one color that can be assigned to that set. Now we confirm that this coloring solves the problem stated at the beginning.

Theorem 3. *C defines a coloring of \mathbf{F} such that no two triangles of the same color share a vertex.*

Proof: Suppose $C(\triangle ABC) = C(\triangle ADE)$. Then $\{L(A), L(B), L(C)\}$ and $\{L(A), L(D), L(E)\}$ are either both in S_n or both in S_n^c , since they are not disjoint. This means that the sets have at least two elements in common. Since all the vertices adjacent to A have distinct labelings (a simple consequence of the construction in Lemma 1), this means that the original triangles also have two points in common. Suppose, without loss of generality, $C = D$. If $B = E$ also, we are done. Otherwise, let $a = L(A)$, $b = L(B)$, $c = L(C)$, $e = L(E)$. By Rule 1, $(a, c, b, e) = -1$. But from Lemma 2 and the fact that $\{a, c, b, e\} = S_n$ or S_n^c , it follows that $(a, c, b, e) = 3$ or 5 . By contradiction, we conclude that $\triangle ABC = \triangle ADE$. ■

Solution 3. We can describe the coloring of Figure 1 in yet another way, using a highlighting of edges instead of a labeling of vertices. Proceeding counterclockwise, highlight the seven central edges of the tessellation in colors $0, 1, \dots, 6$. Extend this to a highlighting of \mathbf{E} by the following rules:

Rule 1. *The three edges of any triangle must be highlighted in colors that correspond to three collinear points in the projective plane illustrated in Figure 3. (Thus, for example, if two edges of a triangle are highlighted 0 and 1, then the third edge must be highlighted 3.)*

Rule 2. *The seven edges incident at any vertex A must be highlighted in colors $\phi(0), \phi(1), \dots, \phi(6)$, proceeding counterclockwise around A , where ϕ is an automorphism of the projective plane. For example, as Figure 4 illustrates, if three consecutive edges are highlighted 0, 6, 5 (proceedings counterclockwise about a point A), then the remaining edges must be highlighted 2, 1, 3, 4, in that order.*

Finally, note that the lines in our projective plane have been labeled in such a way that

$$\{a, b, c\} \text{ are collinear} \Leftrightarrow \{a, b, c, \infty\} = S_n,$$

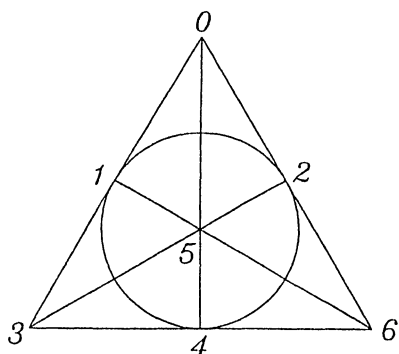


Figure 3. The projective plane of order 7.

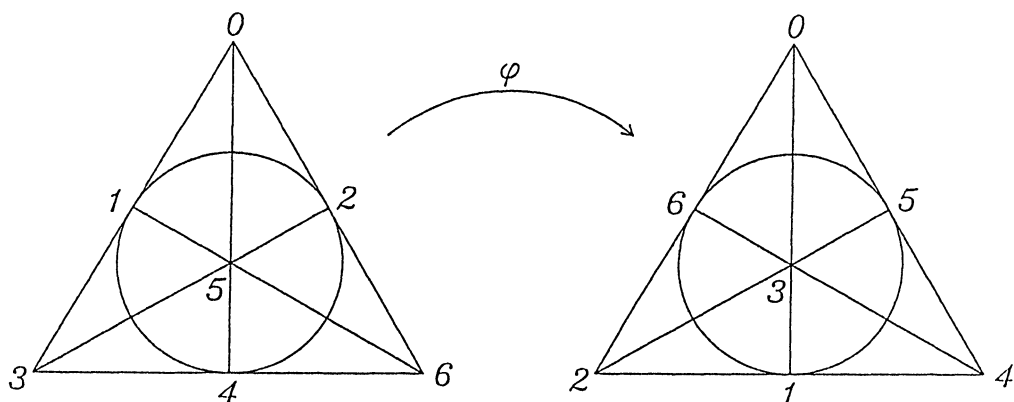


Figure 4. An automorphism of the projective plane in Figure 3.

where S_n is one of the sets defined in Solution 2. This brings us to the final coloring rule:

Rule 3. *If the edges of a triangle are highlighted a, b, c , then color the triangle in color n , where $\{a, b, c, \infty\} = S_n$.*

Though these rules are somewhat clumsier than the rules of Solution 2, the reader can verify that they lead to precisely the same coloring.

THE ISOMORPHISM BETWEEN $PSL_2(\mathbb{Z}_7)$ AND $SL_3(\mathbb{Z}_2)$. It is more or less obvious, from Solution 2, that the symmetry group of the $(2, 3, 7)$ tessellation of the torus T_3 (not the entire hyperbolic plane) “should be” $PSL_2(\mathbb{Z}_7)$. (identify each isometry χ by the labelings of the vertices $\chi(P)$, $\chi(Q)$, $\chi(R)$). Likewise, the symmetry group “should be” $SL_3(\mathbb{Z}_2)$ (which is the symmetry group of the projective plane in Figure 3). Thus a consequence of Solutions 2 and 3 is that $PSL_2(\mathbb{Z}_7)$ and $SL_3(\mathbb{Z}_2)$ are isomorphic. The goal of this section is to construct the isomorphism as explicitly as possible, without the aid of figures. Why? To prove that the identification with $PSL_2(\mathbb{Z}_7)$ is one-to-one and onto depends on examining Figure 2 to make sure that no two triangles have the same vertex labelings and that each

vertex labeling actually occurs. Similarly, the proof of the second identification rests on examining the edges meeting at all 24 vertices to make sure that each automorphism of the projective plane occurs once and no more than once. Though this is easy enough to do, I would contend that it is a very uninformative, “brute force” proof.

Lemma 4. $|PSL_2(\mathbf{Z}_7)| = |SL_3(\mathbf{Z}_2)| = 168$.

Proof: Any element $\chi \in PSL_2(\mathbf{Z}_7)$ is determined uniquely by $\chi(\infty)$, $\chi(0)$, and $\chi(1)$. There are 8 choices for $\chi(\infty)$; 7 choices for $\chi(0)$, and only 3 choices for $\chi(1)$, since only half of the 6 elements of $\mathbf{Z}_7 \cup \{\infty\}$ which are distinct from $\chi(\infty)$ and $\chi(0)$ yield a positively oriented triple. Thus $|PSL_2(\mathbf{Z}_7)| = 8 \cdot 7 \cdot 3 = 168$.

Any element $\chi \in SL_3(\mathbf{Z}_2)$ is also uniquely determined by three values. There are 7 choices for $\chi(1, 0, 0)$ (since $\chi(1, 0, 0) \neq (0, 0, 0)$; 6 choices for $\chi(0, 1, 0)$; and 4 choices for $\chi(0, 0, 1)$. since $\chi(0, 0, 1) \notin \text{Span}\{\chi(1, 0, 0), \chi(0, 1, 0)\}$. Thus $|SL_3(\mathbf{Z}_2)| = 7 \cdot 6 \cdot 4 = 168$. ■

Lemma 5. $SL_3(\mathbf{Z}_2) = \text{Aut}[(\mathbf{Z}_2)^3]$.

Here we are considering $(\mathbf{Z}_2)^3$ only as an additive group and “forgetting” the vector space structure. This lemma is obvious from the fact that any set of three basis vectors of $(\mathbf{Z}_2)^3$ as a vector space is also a set of generators for $(\mathbf{Z}_2)^3$ as a group, and conversely.

We will now define a new operation, \oplus , on $\mathbf{Z}_7 \cup \{\infty\}$, and show that $\langle \mathbf{Z}_7 \cup \{\infty\}, \oplus \rangle$ is isomorphic to $(\mathbf{Z}_2)^3$. In fact, even more is true: $\langle \mathbf{Z}_7 \cup \{\infty\}, \oplus, + \rangle$ is a field, with the ordinary addition (mod 7), $+$, serving as the *multiplication* in the new field! It would be interesting to know whether there exist any other algebraic structures like this, where \cdot distributes over $+$ in \mathbf{Z}_7 , which in turn distributes over \oplus in $\mathbf{Z}_7 \cup \{\infty\}$.

The definition of \oplus is motivated by the coloring procedures given in Solutions 2 and 3 above: if a and b are the labels of two adjacent vertices in Solutions 2, then $a \oplus b$ will be highlighted color of the edge joining them in Solution 3.

Definition 6. If $a \neq 0 \pmod{7}$, let $\chi(a) = (a/7)$, the Legendre symbol of $a \pmod{7}$. (This simplifies some of the typography below.) That is, $\chi(a) = 1$ if a is a quadratic residue modulo 7 (i.e. $a = 1, 2$, or 4), and $\chi(a) = -1$ otherwise.

Definition 7. If $a \neq b \in \mathbf{Z}_7$, then $a \oplus b \equiv c$, where c is the unique element of \mathbf{Z}_7 such that $(a, b, c, \infty) \equiv 5^{\chi(b-a)} \pmod{7}$.

In addition, we define $a \oplus a \equiv \infty$ and $a \oplus \infty = \infty \oplus a \equiv a$. Note that $a \oplus b = b \oplus a$, since $\chi(a - b) = -\chi(b - a)$.

The proof that $\mathbf{Z}_7 \cup \{\infty\}$ is a field rests on two fairly simple lemmas, whose proofs we will leave to the reader.

Lemma 8. Define $\mathcal{S}_n = \{(n, \infty), (n + 1, n + 3), (n + 2, n + 6), (n + 4, n + 5)\}$. Then any ordered pair (a, b) such that $b - a$ is a quadratic residue is an element of a unique set \mathcal{S}_n . Also, for any two distinct ordered pairs, (a, b) and (c, d) in \mathcal{S}_n , $(a, b, c, d) \equiv 5 \pmod{7}$. In particular, it follows that

$$(n + 1) \oplus (n + 3) = n, (n + 2) \oplus (n + 6) = n, \text{ and } (n + 4) \oplus (n + 5) = n.$$

Lemma 9. *If $a, b \in \mathbf{Z}_7$, $a \neq b$, then*

$$a \oplus b = \begin{cases} 5a + 3b & \text{if } \chi(b - a) = 1, \\ 3a + 5b & \text{if } \chi(a - b) = 1. \end{cases} \quad (1)$$

In general, the characterization of \oplus in terms of cross ratios is more useful than equation (1), but Lemma 9 simplifies calculations of \oplus —for example, if we were trying to construct the edge-highlighting of Solution 3 directly from the vertex-labeling of Solution 2. We also need equation (1) in the proof of Theorem 13.

Theorem 10. (i) *If $\lambda \in \mathbf{Z}_7$ then $(a + \lambda) \oplus (b + \lambda) = (a \oplus b) + \lambda$.*

(ii) *If $\chi(\mu) = 1$ then $(\mu a) \oplus (\mu b) = \mu(a \oplus b)$.*

(iii) $\mathbf{Z}_7 \cup \{\infty\}$ is a field, with addition defined by \oplus and multiplication defined by $+$ (mod 7).

Proof: The proofs of (i) and (ii) are obvious from the properties of the cross ratio. Property (ii) is not, of course necessary for $\mathbf{Z}_7 \cup \{\infty\}$ to be a field, but we will use this property below. Note, incidently, that (ii) is false if $\chi(\mu) = -1$.

To prove (iii), it remains to verify the group axioms for \oplus . The only one that is not obvious is associativity. Of course, this could be checked by “brute force,” but again it is more interesting to present a proof which makes use of the properties of the Legendre symbol and cross ratio. Here is a sketch of the argument, with details left to the reader.

First, verify that associativity holds when one term is repeated; that is, for any $a, b \in \mathbf{Z}_7$, $(a \oplus a) \oplus b = a \oplus (a \oplus b) = b$. The fact that $(a \oplus a) \oplus b = b$ is immediate from Definition 7. The second equality follows from the identity $\chi(a \oplus b - a) = \chi(a - b)$, which can be checked using Lemma 9. Next, using Lemma 8 and Lemma 3, verify that for any distinct $a, b, c \in \mathbf{Z}_7$,

$$a \oplus b = c \oplus d \Leftrightarrow (a, b, c, d) \equiv 5^{\chi(b-a)\chi(c-b)\chi(a-c)} \pmod{7}. \quad (2)$$

Finally, using (2), we can prove that associativity holds when all three terms are distinct. This is trivial if one of the terms is ∞ . Otherwise, if a, b, c are distinct elements of \mathbf{Z}_7 ,

$$(a \oplus b) \oplus c = (c \oplus d) \oplus c = d, \text{ where } (a, b, c, d) = 5^{\chi(bb-a)\chi(c-b)\chi(a-c)}$$

and

$$a \oplus (b \oplus c) = a \oplus (a \oplus e) = e, \text{ where } (b, c, a, e) = 5^{\chi(c-b)\chi(a-c)\chi(b-a)}.$$

Hence $(a, b, c, d) = (b, c, a, e) = (a, b, c, e)$, and it follows that $d = e$. ■

Corollary. *As a group, $\mathbf{Z}_7 \cup \{\infty\} \cong (\mathbf{Z}_2)^3$, and hence $\text{Aut}[\mathbf{Z}_7 \cup \{\infty\}] \cong SL_3(\mathbf{Z}_2)$.*

Now we will construct an explicit isomorphism

$$\tau: PSL_2(\mathbf{Z}_7) \rightarrow \text{Aut}[\mathbf{Z}_7 \cup \{\infty\}].$$

The first thing to notice is that we cannot simply let the elements of $PSL_2(\mathbf{Z}_7)$ act on the symbols in $\mathbf{Z}_7 \cup \{\infty\}$ in the normal way, because some linear fractional transformations do not map the identity ∞ to itself, and hence cannot be group automorphisms for the operation \oplus . The correct idea again comes from our colorings of the (2, 3, 7) tessellation.

When an isometry ρ acts on a triangle whose vertices are labeled ∞, a, b , the image (by the construction in Solution 2) will have vertices labeled $[\phi(\rho)](\infty)$, $[\phi(\rho)](a)$, and $[\phi(\rho)](b)$. For simplicity, in Figure 5 we have identified the isometry ρ with $\phi(\rho)$. We claimed above that if a and b are the colors of two

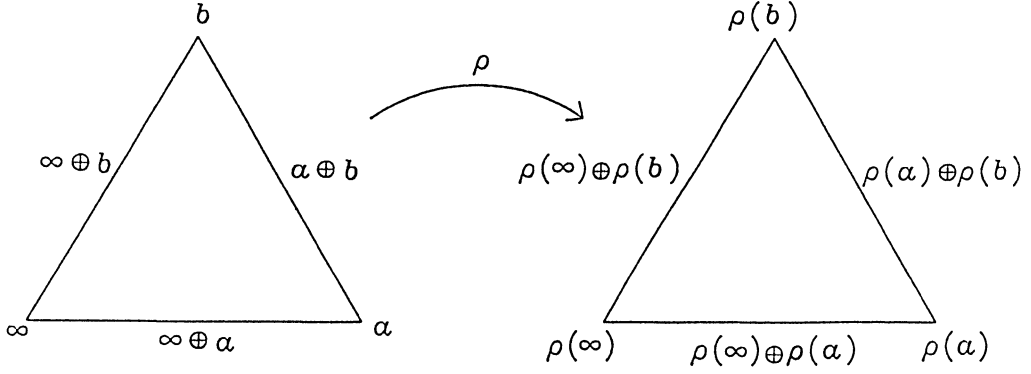


Figure 5. The effect of an isometry ρ on the vertex labeling of Solution 2 and the edge Highlighting of Solution 3.

adjacent vertices in Solution 2, then $a \oplus b$ is the color of the edge joining them in Solution 3. Thus we obtain the edge-highlighting illustrated in Figure 5. In order for the highlighting to be consistent with the action of the isometry ρ , ρ must act in the following way on the edge highlighting:

$$A_\rho(a) = A_\rho(\infty \oplus a) \equiv \rho(\infty) \oplus \rho(a). \quad (3)$$

Then our conjectured isomorphism τ is given by

$$\tau(\rho) = A_\rho. \quad (4)$$

Our task now is twofold: first, to confirm that for any $\rho \in PSL_2(\mathbb{Z}_7)$, the map A_ρ defined in (3) is indeed an automorphism of $\mathbb{Z}_7 \cup \{\infty\}$; second, to confirm that the map τ defined in (4) is an isomorphism.

Lemma 11. *If $\rho(x) = x + b$, $\rho(x) = ax$ (where $\chi(a) = 1$) or $\rho(x) = -x^{-1}$, then A_ρ is an automorphism of $\mathbb{Z}_7 \cup \{\infty\}$.*

Proof: The first two cases follow from Theorem 10. The third case, $\rho(x) = -x^{-1}$, follows from the following computation. For any $x, y \in \mathbb{Z}_7$, such that $x \neq y$:

$$\begin{aligned} (x, y, x \oplus \infty) &= 5^{\chi(y-x)} \\ \Rightarrow (x, y, \infty, x \oplus y) &= 5^{\chi(x-y)} \\ \Rightarrow (-x^{-1}, -y^{-1}, 0, -(x \oplus y)^{-1}) &= 5^{\chi(x-y)} = 5^{\chi(x^{-1}-y^{-1})\chi(-x^{-1})\chi(y^{-1})} \\ \Rightarrow (-x^{-1}) \oplus (-y^{-1}) &= 0 \oplus (-(x \oplus y)^{-1}) \text{ (by equivalence (2))} \\ \Rightarrow A_\rho(x) \oplus A_\rho(y) &= A_\rho(x \oplus y). \end{aligned}$$

The cases where $x = y$ or where x or y equal ∞ are trivial. Likewise, the proof that A_ρ is a bijection is straightforward.

Lemma 12. $A_{\rho_1\rho_2} = A_{\rho_1}A_{\rho_2}$. Hence τ is a homomorphism. Moreover, $A_\rho \in \text{Aut } \mathbb{Z}_7 \cup \{\infty\}$ for all $\rho \in PSL_2(\mathbb{Z}_7)$. ■

Proof: We leave to the reader the verification that $A_{\rho_1\rho_2} = A_{\rho_1}A_{\rho_2}$ if ρ_1 has one of the three forms in Lemma 11. Since these three types of linear fractional transformations generate $PSL_2(\mathbb{Z}_7)$, the same statement holds for all $\rho_1, \rho_2 \in PSL_2(\mathbb{Z}_7)$.

Similarly, since compositions of automorphisms are automorphisms, the second assertion also follows immediately. ■

Theorem 13. τ is an isomorphism. Hence $PSL_2(\mathbf{Z}_7) \cong SL_3(\mathbf{Z}_2)$.

Proof: Even after all this effort, it's still not quite trivial! First we show that τ is injective. Suppose that $A_\phi(x) = x$ for all x . Then $\phi(\infty) \oplus \phi(x) = x$ and $\phi(\infty) \oplus \phi(y) = y$ for all x, y . Adding, we conclude that $\phi(x) \oplus \phi(y) = x \oplus y$ for all x, y . Adding $x \oplus \phi(y)$ to both sides, we have $x \oplus \phi(x) = y \oplus \phi(y)$ for all x, y . Hence there is a constant A such that $x \oplus \phi(x) = A$ for all x , or $\phi(x) = x \oplus A$. We claim this cannot be a linear fractional transformation unless $A = \infty$. Otherwise, from Lemma 9 we have

$$\phi(x) = \begin{cases} 5x + 3A & \text{if } \chi(A - x) = 1 \\ 3x + 5A & \text{if } \chi(A - x) = -1. \end{cases}$$

But since $\phi(x)$ agrees with the linear fractional transformation $5x + 3A$ for the three values of x such that $\chi(A - x) = 1$, and since linear transformations are determined by their values at three points, then $\phi(x) = 5x + 3A$ for all x . This contradicts the equation above. Hence $A = \infty$, and $\phi(x) = x \oplus \infty = x$, so that ϕ is the identity map.

The very last step is trivial. Since τ is one-to-one, and by Lemma 4 the number of elements in its domain and range are equal, τ is onto. Hence τ is an isomorphism. ■

FINAL REMARK. We have seen that the 7-coloring of the hyperbolic plane in Figure 1 is intimately connected with the existence of an isomorphism between PSL_2 and $SL_3(\mathbf{Z}_2)$. But this isomorphism is highly unusual: in fact, it is the *only* isomorphism between finite simple groups of Lie type listed in [2]. It would be interesting to investigate whether d -colorings of the $(2, 3, d)$ tessellation for $d > 7$ exist; whether they can be associated with vertex and edge colorings as we have done here; and, if so, whether any interesting group-theoretic consequences may follow.

REFERENCES

1. Cole, K. C. Escape from 3-D. *Discovery*, July 1993, 52–62.
2. Conway, J., et al., eds *Atlas of Finite Groups*. Oxford University Press, Oxford, 1984.
3. Stillwell, J. *Geometry of Surfaces*. Springer-Verlag, New York, 1992.
4. Thompson, T. M. *From Error-Correcting Codes Through Sphere Packings to Simple Groups*. Mathematical Association of America, Washington, 1983.
5. *Notices of the American Mathematical Society* 41 (1994), 31–32.

Department of Mathematics
Kenyon College
Gambier, OH 43022
mackenzi@kenyon.edu

Cosine Products, Fourier Transforms, and Random Sums

Kent E. Morrison

1. INTRODUCTION. The function $\sin x/x$ is endlessly fascinating. By setting $x = \pi/2$ in the infinite product expansion

$$\frac{\sin x}{x} = \prod_{k=1}^{\infty} \cos \frac{x}{2^k} \quad (1)$$

one gets the first actual formula for π that mankind ever discovered, dating from 1593 and due to François Viète (1540–1603), whose Latinized name is Vieta. (Was any notice taken of the formula’s 400th anniversary, perhaps by the issue of a postage stamp?) From the samples of a function $f(x)$ at equally spaced points x_n , $n \in \mathbf{Z}$, one can reconstruct the complete function with the aid of $\sin x/x$, provided f is “band-limited” and the spacing of the samples is small enough. This is the content of the Sampling Theorem, which lends its name to $\sin x/x$ as the **sampling function**. Its importance in signal processing, where it is also known as $\text{sinc } x$, is the result of its Fourier transform being the characteristic function of the interval $[-1, 1]$ (modulo a scalar factor).

In Section 2 we prove the infinite product expansion for $\sin x/x$ and derive Viète’s formula. In Section 3 we transform the product expansion with the Fourier transform and use convolution and delta distributions to prove it in a way that reveals a host of similar identities. Section 4 puts these identities into a probabilistic setting, and in Section 5 we alter the probability experiments in order to make connections between infinite cosine products, Cantor sets, and sums of series with random signs, particularly the harmonic series. This leaves us with some interesting unsolved problems and conjectures for further work.

2. AN ELEMENTARY PROOF. Repeated use of the double angle formula for the sine shows that

$$\begin{aligned} \sin x &= 2 \sin \frac{x}{2} \cos \frac{x}{2} \\ &= 4 \sin \frac{x}{4} \cos \frac{x}{4} \cos \frac{x}{2} \\ &\vdots \\ &= 2^n \sin \frac{x}{2^n} \left(\prod_{k=1}^n \cos \frac{x}{2^k} \right). \end{aligned}$$

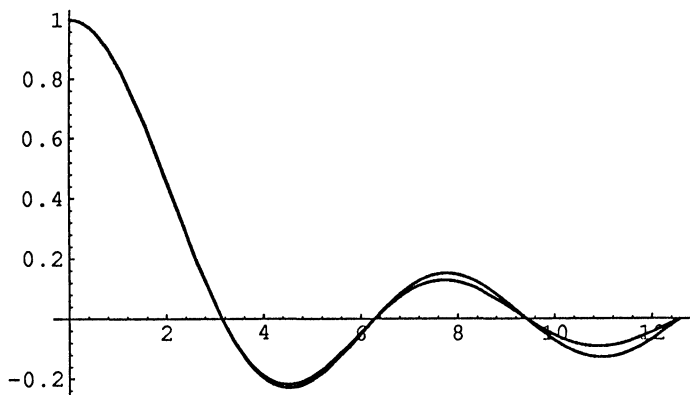


Figure 1. Graphs of $\sin x/x$ and $\cos x/2 \cos x/4 \cos x/8$. Where both graphs are visible, $\sin x/x$ is nearer the x axis.

But

$$\lim_{n \rightarrow \infty} 2^n \sin \frac{x}{2^n} = x,$$

thereby proving the identity. See Figure 1 for an indication of how quickly the product converges.

Let $x = \pi/2$, make use of the half-angle identity, and there you have Viète's formula for π ,

$$\frac{2}{\pi} = \frac{\sqrt{2}}{2} \frac{\sqrt{2 + \sqrt{2}}}{2} \frac{\sqrt{2 + \sqrt{2 + \sqrt{2}}}}{2} \dots \quad (2)$$

At this point the cosine identity could remain an isolated curiosity of historical interest, relegated to the ends of exercise sets in textbooks. In fact, it is just the first of an infinite family of cosine product identities for $\sin x/x$.

3. THE FOURIER TRANSFORM AND MORE IDENTITIES. For a complex valued function $f(x)$ defined on the real line, the Fourier transform puts together f as a continuous linear combination of the "pure" oscillations $e^{i\omega x}$ in which the coefficient in front of $e^{i\omega x}$ is denoted by $\hat{f}(\omega)$. Thus,

$$f(x) = \int_{-\infty}^{\infty} \hat{f}(\omega) e^{i\omega x} d\omega. \quad (3)$$

The function \hat{f} is the **Fourier transform** of f and the integral above is a description of how to get back f from \hat{f} and is actually the formula for the inverse transform. How do we get \hat{f} from f ? That is given by this integral:

$$\hat{f}(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(x) e^{-i\omega x} dx. \quad (4)$$

Of course, the proofs of these relationships involve hypotheses on the functions so that the integrals make sense, but they can be extended beyond the realm of ordinary functions to generalized functions or distributions. We need more than ordinary functions in order to make sense of the Fourier transform of a sine or cosine.

Notation: we also write the Fourier transform of f as $\mathcal{F}(f)$ and the inverse transform of ϕ as $\mathcal{F}^{-1}(\phi)$.

Consider $\cos bx$, which by Euler's Identity may be written as

$$\cos bx = \frac{1}{2}(e^{ibx} + e^{-ibx}).$$

This shows the function written as a linear combination of just two of the functions $e^{i\omega x}$ for $\omega = b$ and $\omega = -b$. The coefficients appear to be $1/2$, but if we use them in the integral form with all other coefficients zero, then we cannot represent the cosine function. Instead, we must regard the coefficients as point masses at b and $-b$. Therefore, the Fourier transform of $\cos bx$ is $(1/2)(\delta_b + \delta_{-b})$, where δ_b denotes the Dirac delta distribution or point mass at the point b . All of this can be made rigorous, but at the expense of some long development in graduate level analysis. The approach here is at about the level of a second year course in engineering mathematics.

In addition, the Fourier transform behaves nicely on a product of functions and turns it into the convolution of the transforms:

$$\widehat{fg} = \hat{f} * \hat{g}. \quad (5)$$

For two functions $\phi(\omega)$ and $\psi(\omega)$, the convolution $\phi * \psi$ is defined by

$$(\phi * \psi)(\omega) = \int_{-\infty}^{\infty} \phi(\alpha) \psi(\omega - \alpha) d\alpha. \quad (6)$$

Again, we must extend convolution beyond the realm of functions. In particular we need convolutions of delta distributions and for them we can easily show that δ_0 behaves as the identity for convolution

$$\delta_0 * \phi = \phi \quad (7)$$

and that

$$\delta_a * \delta_b = \delta_{a+b}. \quad (8)$$

Now back to the cosine identity. Let $f(x) = \prod_{k=1}^{\infty} \cos(x/2^k)$ and let f_n be the n th partial product. The Fourier transform of f_n is

$$\hat{f}_n = * \prod_{k=1}^n \frac{1}{2} (\delta_{1/2^k} + \delta_{-1/2^k}).$$

The asterisk in front of the product sign indicates a repeated convolution of the factors. Expanding for $n = 3$ we see that

$$\hat{f}_3 = \frac{1}{8} (\delta_{-7/8} + \delta_{-5/8} + \cdots \delta_{7/8}).$$

Likewise

$$\hat{f}_n = \frac{1}{2^n} \sum_{b \in B_n} \delta_b,$$

where B_n is the set of 2^n equally spaced numbers from $-1 + 1/2^n$ to $1 - 1/2^n$ with spacing $2/2^n = 1/2^{(n-1)}$.

The sequence of measures \hat{f}_n converges to the uniform density on $[-1, 1]$ of total mass 1, which we can write as $(1/2)\chi_{[-1, 1]} d\omega$. The inverse transform is easy to compute:

$$\int_{-1}^1 \frac{1}{2} e^{i\omega x} d\omega = \frac{\sin x}{x}.$$

The spectrum of $(\sin x)/x$ is uniform in the interval $-1 \leq \omega \leq 1$. This means that $\sin x/x$ is a continuous linear combination of the “pure” harmonics $e^{i\omega x}$ with the same weight of $1/2$ for each $\omega \in [-1, 1]$.

With this proof we have a way to generate a family of similar identities. Let us put point masses at 3^n equally spaced points from $-1 + 1/3^n$ to $1 - 1/3^n$ with spacing $2/3^n$. Such a measure is the convolution $*\prod_{k=1}^n \frac{1}{3}(\delta_{-2/3^k} - \delta_0 + \delta_{2/3^k})$. Applying the inverse transform

$$\mathcal{F}^{-1}\left(\frac{1}{3}(\delta_{-2/3^k} + \delta_0 + \delta_{2/3^k})\right) = \frac{1}{3}\left(2\cos\frac{2x}{3^k} + 1\right)$$

and taking limits gives us the infinite product identity

$$\prod_{n=1}^{\infty} \frac{1}{3}\left(1 + 2\cos\frac{2x}{3^n}\right) = \frac{\sin x}{x}. \quad (9)$$

Let us use the positive integer p as the base (we have just seen $p = 2$ and $p = 3$). The first measure \hat{f}_1 is the sum of point masses at p points equally spaced from $-1 + 1/p$ to $1 - 1/p$ with spacing $2/p$.

$$\hat{f}_1 = \frac{1}{p}\left(\delta_{\frac{1-p}{p}} + \delta_{\frac{3-p}{p}} + \delta_{\frac{5-p}{p}} + \cdots + \delta_{\frac{p-1}{p}}\right) \quad (10)$$

$$= \frac{1}{p} \sum_{l=0}^{p-1} \delta_{\frac{2l+1-p}{p}} \quad (11)$$

We let

$$\hat{f}_n = * \prod_{k=1}^n \left(\frac{1}{p} \sum_{l=0}^{p-1} \delta_{\frac{2l+1-p}{p^k}} \right)$$

and one can see that \hat{f}_n consists of p^n point masses equally spaced from $-1 + 1/p^n$ to $1 - 1/p^n$ with spacing $2/p^n$. Taking the inverse transform we see that

$$\mathcal{F}^{-1}(\hat{f}_n)(x) = \prod_{k=1}^n \frac{1}{p} \sum_{l=0}^{p-1} \exp((2l+1-p)ix/p^k).$$

Rewriting the exponential as cosines and taking limits gives the general identities.

There is a slight difference in the form depending on the parity of p . For p even

$$\prod_{k=1}^{\infty} \frac{1}{p} \left(\sum_{\substack{1 \leq m \leq p-1 \\ m \text{ odd}}} 2\cos\frac{mx}{p^k} \right) = \frac{\sin x}{x}. \quad (12)$$

For p odd

$$\prod_{k=1}^{\infty} \frac{1}{p} \left(1 + \sum_{\substack{1 \leq m \leq p-1 \\ m \text{ even}}} 2\cos\frac{mx}{p^k} \right) = \frac{\sin x}{x}. \quad (13)$$

For $p = 6$ the identity takes the form

$$\prod_{k=1}^{\infty} \frac{1}{6} \left(2\cos\frac{x}{6^k} + 2\cos\frac{3x}{6^k} + 2\cos\frac{5x}{6^k} \right) = \frac{\sin x}{x}. \quad (14)$$

For $p = 7$ the identity takes the form

$$\prod_{k=1}^{\infty} \frac{1}{7} \left(1 + 2 \cos \frac{2x}{7^k} + 2 \cos \frac{4x}{7^k} + 2 \cos \frac{6x}{7^k} \right) = \frac{\sin x}{x}. \quad (15)$$

For larger p fewer terms in the product are needed for the same degree of accuracy in the approximation to $\sin x/x$. In fact, by letting p go to infinity the first factor alone approaches $\sin x/x$ and provides a novel derivation of a well-known result. I leave it to the reader to work it out.

4. PROBABILISTIC INTERPRETATION. Mark Kac, in his delightful and now classic Carus monograph [2], proves the first cosine identity (1) in a way that is equivalent to the one we have outlined, although he does not explicitly use the Fourier transform, delta functions, and convolution. He then turns the identity into a question of probability, which for him was the leitmotif of his mathematical work.

The original product identity (1) arises from the following experiment. Flip a fair coin repeatedly. Beginning with 0, add $1/2$ if the result is heads and subtract $1/2$ if the result is tails. On the next toss add or subtract $1/4$; on the next add or subtract $1/8$, and so on. What is the distribution of the sums over the probability space whose elements are the countable sequences of coin tosses? Clearly the sums are distributed uniformly between -1 and 1 .

Let s_n denote the n th partial sum. It is a sum of independent random variables $a_1 + a_2 + \cdots + a_n$, where a_k has the probability distribution $(1/2)(\delta_{1/2^k} + \delta_{-1/2^k})$. The probability distribution of a sum of independent random variables is the convolution of the respective distributions of the random variables. Therefore, s_n has the distribution

$$* \prod_{k=1}^n \frac{1}{2} (\delta_{1/2^k} + \delta_{-1/2^k}) = \frac{1}{2^n} \left(\delta_{\frac{1-2^n}{2^n}} + \cdots + \delta_{\frac{2^n-1}{2^n}} \right).$$

The inverse Fourier transform of a probability measure is called its **characteristic function**. Thus, the characteristic function for the distribution of s_n is the product $\prod_{k=1}^n \cos x/2^k$. In the theory of probability and statistics, characteristic functions are a powerful tool. Typically computations are done with characteristic functions in order to draw conclusions about distributions of random variables as in the standard proof of the Central Limit Theorem. Here, however, we have inverted the relationship in order to compute with the probability measures and to get results about the characteristic functions.

5. RELATED PRODUCTS: EXAMPLES AND CONJECTURES

5.1. Coin tossing and Cantor sets. The Cantor set K is the set of points between 0 and 1 whose ternary expansion has no 1's in it. So z is in K if $z = \sum_{k=1}^{\infty} t_k 3^{-k}$, $t_k \in \{0, 2\}$. Define K_n to be the set of elements of K that have the form $\sum_{k=1}^n t_k 3^{-k}$, and define a probability measure supported on K_n

$$\mu_n = \frac{1}{2^n} \sum_{z \in K_n} \delta_z. \quad (16)$$

K_n has 2^n elements so μ_n is equally distributed on K_n . The sequence (μ_n) has a limit μ , which can be described as assigning the following limit as the measure of a set E :

$$\mu(E) = \lim_{n \rightarrow \infty} \frac{\#E \cap K_n}{2^n}. \quad (17)$$

The measure μ is also the Lebesgue-Stieltjes measure of the Cantor function. The Cantor function is continuous, non-decreasing, and has derivative zero on the complement of the Cantor set. Thus it defines a measure supported on the Cantor set, which is precisely the measure μ defined in (17).

What is of interest in this note is that μ_n is the finite convolution product

$$\mu_n = * \prod_{k=1}^n \frac{1}{2} (\delta_0 + \delta_{2/3^k}). \quad (18)$$

Consider the experiment of tossing a fair coin. On toss number k let

$$a_k = \begin{cases} 0 & \text{heads} \\ 2/3^k & \text{tails} \end{cases}.$$

Let $s_n = \sum_{k=1}^n a_k$. Then s_n is equally distributed over K_n . The characteristic function for the distribution of s_n is $\prod_{k=1}^n (1/2)(1 + e^{2xi/3^k})$. Define

$$f(x) = \prod_{k=1}^{\infty} \frac{1}{2} (1 + e^{2xi/3^k}). \quad (19)$$

(One checks easily that the product is convergent.) Then $\hat{f} = \mu$, the Cantor measure, but is it possible to characterize f in any other way?

This leads us to look at the related infinite product $\prod_{k=1}^{\infty} \cos 2x/3^k$. Because

$$\mathcal{J}\left(\cos \frac{2x}{3^k}\right) = \frac{1}{2} (\delta_{2/3^k} + \delta_{-2/3^k})$$

the probabilistic interpretation is clear: add or subtract $2/3^k$ on the k th toss with equal probability. Let s_n be the sum of the first n values. What is the distribution of s_n and what is the distribution of $s = \lim_{n \rightarrow \infty} s_n$? The exercise of expanding and plotting the values of s_3 lead one to suspect that s is distributed “uniformly” over the Cantor set constructed from $[-1, 1]$ by successively removing middle thirds. That is easy to prove, as follows.

Define the affine map of $[0, 1]$ to $[-1, 1]$ by $z \mapsto 2(z - 1/2)$. Let $z = \sum t_k 3^{-k}$, $t_k \in \{0, 2\}$, be a point in the Cantor set. The ternary expansion of $1/2$ is $\sum 3^{-k}$, and so $2(z - 1/2) = \sum 2(t_k - 1)3^{-k}$. The coefficients $2(t_k - 1)$ are either 2 or -2 with equal probability.

This shows that the infinite product $\prod_{k=1}^{\infty} \cos 2x/3^k$ has Fourier transform equal to the Cantor measure on the Cantor set constructed from $[-1, 1]$ by removing middle thirds, but it does not give us a closed form like $\sin x/x$. It would be most surprising if there were any simpler description of $\prod_{k=1}^{\infty} \cos 2x/3^k$. In Figure 2 is a plot of the partial product with $n = 8$ and $0 \leq x \leq 100$. (The function is even.) Over this range the infinite product is indistinguishable from the eighth partial product. The self-similarity of the Cantor set at smaller and smaller scales appears to be reflected in the self-similarity of the graph at higher and higher frequencies.

5.2. Harmonic Series with Random Signs. We have been looking at the sums of series of the form

$$\sum_{k=1}^{\infty} t_k c_k \quad (20)$$

where t_k is randomly chosen to be 1 or -1 with equal probability. Rademacher proved that if $\sum c_k^2 < \infty$, then the sum converges with probability one on the probability space $\Omega = \{-1, 1\}^{\mathbb{N}}$. (Ω can be identified with the unit interval and the

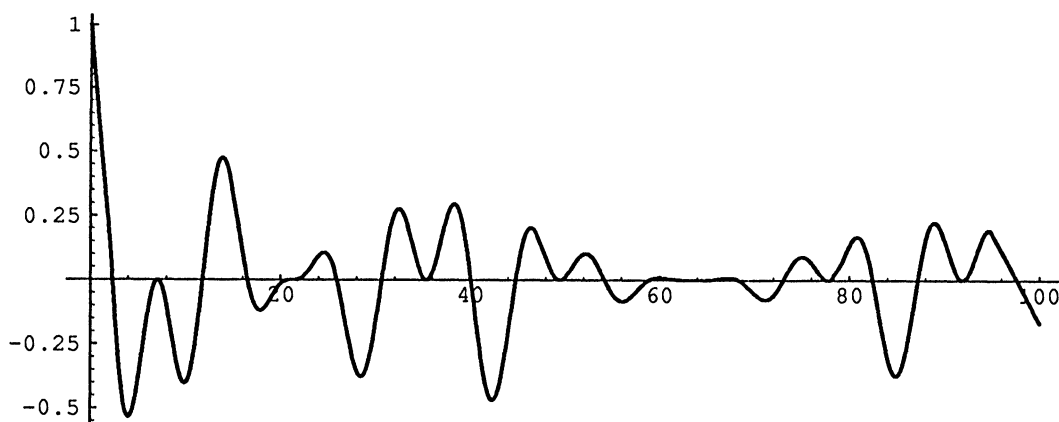


Figure 2. Graph of $\prod_{k=1}^8 \cos \frac{2x}{3^k}$.

probability measure with Lebesgue measure by using binary representations of number: in the interval.) In [2] Kac gives the proof of this theorem due to Paley and Zygmund. It is also a theorem that the series diverges with probability one if $\sum c_k^2 = \infty$. Let us consider the random harmonic series

$$\sum_{k=1}^{\infty} \frac{t_k}{k}, \quad (21)$$

which converges almost surely by Rademacher's result, with the goal of understanding the distribution of the sums. This means we want to understand the distribution of the random variable s defined on Ω . If we let s_n be the partial sum, also a random variable, then the probability distribution of s_n is the measure

$$\mu_n = * \prod_{k=1}^n \frac{1}{2} (\delta_{1/k} + \delta_{-1/k}) \quad (22)$$

and its inverse transform is

$$\mathcal{F}^{-1}(\mu)(x) = \prod_{k=1}^n \cos \frac{x}{k}. \quad (23)$$

The product converges uniformly on compact sets as $n \rightarrow \infty$, and so it is plausible that the sequence μ_n converges to a probability measure μ that is the distribution of the random variable s . There is, however, a fair bit of analysis to make this rigorous. Assuming that the analysis can be made rigorous, then the plot of the Fourier transform of the infinite product $\prod_{k=1}^{\infty} \cos x/k$ will show how the sums are distributed. Let us call this function $\phi(\omega)$. Then

$$\phi(\omega) = \mathcal{F}\left(\prod_{k=1}^{\infty} \cos \frac{x}{k}\right)(\omega) \quad (24)$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-i\omega x} \prod_{k=1}^{\infty} \cos \frac{x}{k} dx \quad (25)$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} (\cos \omega x + i \sin \omega x) \prod_{k=1}^{\infty} \cos \frac{x}{k} dx \quad (26)$$

$$= \frac{1}{\pi} \int_0^{\infty} \cos \omega x \prod_{k=1}^{\infty} \cos \frac{x}{k} dx. \quad (27)$$

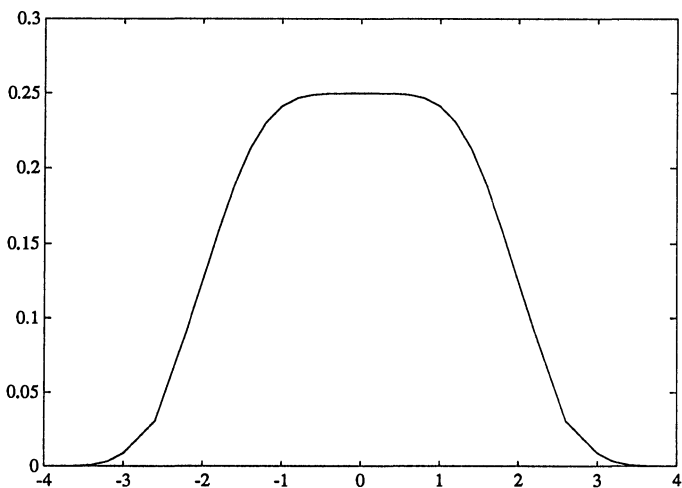


Figure 3. Graph of $\phi(\omega)$.

There is not a closed form for $\phi(\omega)$ and so we resort to numerical integration. We truncated the infinite product at $n = 1000$ and integrated from 0 to 15 using a straightforward Riemann sum with $dx = 0.02$ and the midpoints of the subintervals for the points of evaluation. Values for ω were from 0 to 3.8 in multiples of 0.2. The integration was done with True BASIC on a portable Macintosh. See Figure 3. The distribution is very flat for $-1 < \omega < 1$, much flatter than a normal distribution. A few of the computed values are given in this table. The value of $\phi(0)$ is suspiciously close to $1/4$, suggesting perhaps that $\pi/4$ is the value of the integral

$$\int_0^\infty \prod_{k=1}^\infty \cos \frac{x}{k} dx. \quad (28)$$

One might also conjecture that $\int_0^\infty \cos 2x \prod_{k=1}^\infty \cos x/k dx = \pi/8$.

For additional evidence we turned to simulations of the sums. Using MATLAB we ran 5000 sums of $\sum_{k=1}^{100} t_k/k$ with the values of t_k picked randomly as ± 1 with equal probability. Figure 4 shows a histogram of the sums.

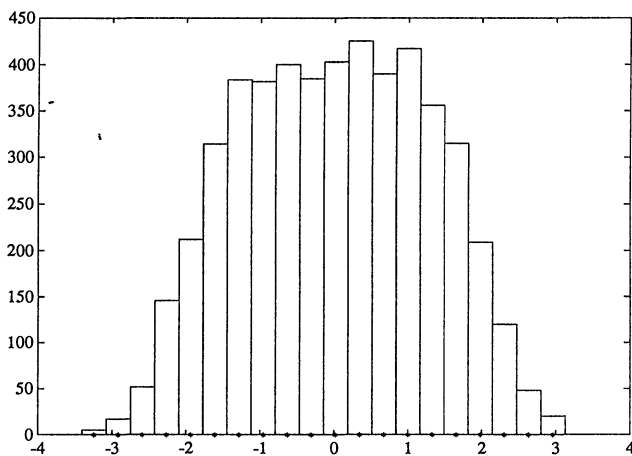


Figure 4. Histogram of 5000 random sums.

ω	$\phi(\omega)$
0.0	.249995
0.1	.249991
0.2	.249972
0.4	.249809
0.6	.249092
0.8	.246819
1.0	.241289
1.2	.230494
1.4	.212941
1.6	.188425
1.8	.158271
2.0	.125000
2.2	.091729
2.4	.061576
2.6	.030596
2.8	.019506
3.0	.008711
3.2	.003181
3.4	.000908
3.6	.000192
3.8	.000028

REFERENCES

1. W. B. Gearhart and H. S. Schultz. The Function $\sin x/x$. *The College Mathematics Journal*, 21:90–99, 1990.
2. M. Kac. *Statistical Independence in Probability, Analysis and Number Theory*. Carus Monographs, no. 12. Mathematical Association of America, Washington, D.C., 1959.

Department of Mathematics
California Polytechnic State University
San Luis Obispo, CA 93407
kmorriso@oboe.calpoly.edu

Proof requires a person who can give and a person who can receive . . .

—*Augustus De Morgan (1808–1871)*

Budget of Paradoxes. London: 1872, p. 262.

How to Add Fast—on Average

Geza Schay

In a recent article in this Monthly C. C. McGeoch [1] described an ingenious method for parallel addition of two n -bit binary integers in $2\log_2 n + 1$ steps, as well as related methods for the sum of three or more integers. These constructions have suggested the subject of the present paper, a very simple, iterative algorithm for the sum of two numbers, which works on average about twice as fast, that is, in about $\log_2 n$ steps. Thus it adds two 64-bit numbers in about six short steps. This is close to amazing, since these are numbers on the order 10^{19} . Apparently, however, the variable length of the method limits its practical usefulness, but it leads to the interesting mathematical problem of finding the probability distribution and the expected value of the length.

This length, that is, the number of steps in the new algorithm, is determined by the length of the longest “carry sequence” (i.e. the longest string of consecutive non-zero carries) in the ordinary addition algorithm as applied to binary numbers. In a classic paper Burks, Goldstine and von Neumann [2, pp. 45–46] have discussed the latter length and obtained $\log_2 n$ as an upper bound for its expected value. However, they did not propose any new algorithm to take advantage of this bound; they only suggested that “either the carries must be accelerated, or use must be made of the average number of carries or both.” Our algorithm, which does this, has been described in the book by Scott [3, pp. 54–55], although without any discussion of the number of steps. In the present paper we make the connection between the new algorithm and the mathematical analysis of [2], and carry it somewhat further by providing an asymptotic formula rather than an upper bound for the expected value, and an approximate evaluation of the probability distribution as well.

THE ALGORITHM. We start with an example. We add the binary digits without any carries and in a second number we save the carries, including zeroes, in the places where the carries would go. We repeat this until all the carry digits become zero:

$$\begin{array}{r}
 \begin{array}{cccccccc}
 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\
 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \\
 \hline
 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\
 \hline
 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
 \hline
 1 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
 \end{array}
 \end{array}$$

In general, let $a = a_n a_{n-1} \dots a_0$ and $b = b_n b_{n-1} \dots b_0$ be two nonnegative integers in binary form and define recursively, for $i = 0, \dots, n$ and $k = 1, 2, \dots$,

$$a_i(0) = a_i, \quad b_i(0) = b_i, \quad (1)$$

$$a_i(k) = a_i(k-1) + b_i(k-1) \bmod 2, \quad (2)$$

$$b_{i+1}(k) = a_i(k-1) \cdot b_i(k-1) \text{ and } b_0(k) = 0. \quad (3)$$

The computation ends when we reach the first k such that $b_i(k) = 0$ for every i , in which case the $a_i(k)$, for $i = 0, \dots, n+1$, constitute the digits of $a + b$.

In the example the algorithm has thus produced the sum $a + b = 1011010100$ in three steps. Clearly it always leads to the sum, since the carries are added in at the same places as in conventional addition, they are just handled in a different order.

THE NUMBER OF STEPS. In the worst case we need $n + 2$ steps. For randomly chosen addends, however, the worst case almost never happens and Table 1 near the end of the paper will indicate, in addition to other results, that the number of steps is under $\log_2 n + 4$ in more than 99% of the cases. The example above is rather typical in this respect.

While a proof for $\log_2 n$ as an approximation to the expected number of steps is fairly involved, it is easy to give a heuristic argument for this expectation: For any value k in the computation, if the carry digit $b_i(k)$ is 1, then $b_{i+1}(k+1)$ is 1 if and only if $a_i(k)$ is 1 as well. Since, by the assumed randomness, the latter event occurs with probability $1/2$, we can expect on average about half as many 1's among the $b_i(k+1)$ as among the $b_i(k)$. Thus, if $n = 2^k$, then half this number of 1's can be halved about $k = \log_2 n$ times to make all carry digits zero.

Let us turn now to the detailed discussion of the probabilities involved.

First, we assume that the digits of a and b are independent Bernoulli random variables, each digit being 0 or 1 with probability $1/2$.

Examining the algorithm we see that in the first step a carry digit 1 is generated only when corresponding digits a_i and b_i are both 1. (In the example above this occurs for $i = 0, 3, 5, 8$.) In the next step, such a carry digit is propagated to $k = 2$ if and only if it is added to a 1. (In the example this occurs only for $i = 1$.) Such a 1, however, can come only from a combination 0, 1 or 1, 0 for a_i and b_i . If a_i and b_i are both 0 or both 1, then their sum is $0 \bmod 2$, and so the propagation of the carry digit 1 is stopped. (In the example this happens with the carry digits $b_4(1) = b_6(1) = b_9(1) = 1$.) The propagation of the carries follows the same rules for all values of k . Thus we see that a carry digit 1 is started at $k = 0$ at combinations of the form 1, 1. Such carry digits are propagated to higher values of k by adjacent combinations at $k = 0$ of the form 0, 1 or 1, 0, and are stopped by combinations at $k = 0$ of the form 0, 0 or 1, 1.

For example in the addition of the two numbers below

$$\begin{array}{cccccccccc} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{array}$$

the 1, 1 at $i = 0$ would generate a 1-valued carry, which would then be propagated by the 1, 0 and the 0, 1, and finally stopped by the 0, 0 at $i = 3$, that is, would result in 0-valued carries up to $i = 3$ when $k = 4$. At the same time the 1-valued carry generated by the 1, 1 at $i = 4$ would be immediately stopped by the 0, 0 at $i = 5$. The same would happen to the carry from $i = 9$, while the carry generated at $i = 6$ would be stopped by the 1, 1 at $i = 9$. Thus all carry digits would become 0 when $k = 4$.

The considerations above suggest the following definition: Given two nonnegative integers $a = a_n a_{n-1} \dots a_0$ and $b = b_n b_{n-1} \dots b_0$ in binary form, we shall call a

sequence of consecutive pairs (a_i, b_i) a *run of length r* if it starts (from the right) by $(1, 1)$, is followed by $r - 1$ pairs of the form $(0, 1)$ or $(1, 0)$, and ends with $(0, 0)$, $(1, 1)$, or with (a_n, b_n) if the latter is the last of the $r - 1$ pairs of the form $(0, 1)$ or $(1, 0)$.

Thus the two numbers in the first example have a run of length 2 and three runs of length 1, whereas the two numbers of the second example have two runs of length 3 and two runs of length 1.

Notice the exceptional case in the definition, when the run ends at the left end. For instance the two examples below both show runs of length 2 (and the second one also a run of length 1):

$$\begin{array}{cc} 0 & 1 & & 1 & 0 & 1 \\ & 1 & 1, & & 1 & 1 & 1. \end{array}$$

(One could, of course, modify the numbers in the first case by starting them with leading zeroes, but that would create a problem in our recursion below.)

It follows at once from the foregoing that the number of steps equals one plus the length of the longest run. Thus we want to obtain the distribution of this length.

Define the following random variables: For the random pair $a = a_n a_{n-1} \dots a_0$ and $b = b_n b_{n-1} \dots b_0$ let $L = \text{length of last (that is, left-most) run}$, $M_n = \text{length of longest run}$. We are interested in finding the distribution of M_n . The simplest way to do this is by writing a recursive formula for the tail probabilities of M_n as follows. The event $\{M_n > r\}$, for any $r = 0, \dots, n$, can be decomposed as the union of two mutually exclusive events: $\{M_{n-1} > r\}$ and the event $\{M_n > r, M_{n-1} \leq r\}$. The latter event can occur if and only if the last run has length $r + 1$ and the longest run in the remaining places has length r or less. Thus we can write for the corresponding probabilities

$$P(M_n > r) = P(M_{n-1} > r) + P(L = r + 1, M_{n-1} \leq r). \quad (4)$$

Writing $q_{nr} = P(M_n > r)$ and making use of the known special form of the last run and the independence of the digits, we obtain the difference equation

$$q_{nr} = q_{n-1,r} + \frac{1}{4} \left(\frac{1}{2} \right)^r (1 - q_{n-r-1,r}). \quad (5)$$

Clearly, we also have $q_{nr} = 0$ if $r > n$.

The above equations can be solved by the method of generating functions and partial fractions. (See Feller [4].) Multiplying through by x^n and summing over n , we get

$$\sum_{n=r}^{\infty} q_{nr} x^n = x \sum_{n=r}^{\infty} q_{n-1,r} x^{n-1} + \sum_{n=r}^{\infty} \frac{1}{4} \left(\frac{1}{2} \right)^r x^n - \frac{x}{4} \left(\frac{x}{2} \right)^r \sum_{n=r+1}^{\infty} q_{n-r-1,r} x^{n-r-1}. \quad (6)$$

Denoting the generating function by $Q_r(x)$ and summing the geometric series in the second term on the right, we can rewrite this as

$$Q_r(x) = xQ_r(x) + \frac{1}{4} \left(\frac{x}{2} \right)^r \frac{1}{1-x} - \frac{x}{4} \left(\frac{x}{2} \right)^r Q_r(x). \quad (7)$$

Hence

$$Q_r(x) = \frac{x^r}{(1-x)(x^{r+1} - 2^{r+2}x + 2^{r+2})}. \quad (8)$$

The probabilities we want to find are the coefficients in the power series expansion of this $Q_r(x)$. To compute them we express the right hand side as a sum of partial fractions and expand those as geometric series. It is easy to see that the second factor in the denominator has a simple zero at some x_1 near 1. This x_1 can be found approximately by setting $x = 1 + z$ in the equation

$$x^{r+1} - 2^{r+2}x + 2^{r+2} = 0 \quad (9)$$

and neglecting $o(z)$ terms. This results in the equation

$$1 + (r + 1)z - 2^{r+2}z \approx 0 \quad (10)$$

and so

$$z \approx \frac{1}{2^{r+2} - r - 1} \quad \text{and} \quad x_1 \approx 1 + \frac{1}{2^{r+2} - r - 1}. \quad (11)$$

Notice that every zero of the denominator in Equation (8) is simple. For if we write

$$f(x) = x^{r+1} - 2^{r+2}x + 2^{r+2}, \quad (12)$$

then any multiple zero \bar{x} of this polynomial would also satisfy

$$f'(\bar{x}) = (r + 1)\bar{x}^r - 2^{r+2} = 0, \quad (13)$$

and so

$$\bar{x}^r = \frac{2^{r+2}}{r + 1}. \quad (14)$$

Substituting this into $f(\bar{x}) = 0$ we would obtain

$$\left(\frac{2^{r+2}}{r + 1} - 2^{r+2} \right) \bar{x} + 2^{r+2} = 0 \quad (15)$$

with the solution

$$\bar{x} = \frac{r + 1}{r}. \quad (16)$$

Equations (14) and (16) are incompatible for integer values of r and so no multiple zeroes can exist.

Denoting the zeroes of $f(x)$ by x_1, x_2, \dots, x_{r+1} we can thus decompose $Q_r(x)$ into partial fractions as

$$Q_r(x) = x^r \left[\frac{A_0}{1 - x} + \frac{A_1}{x_1 - x} + \frac{A_2}{x_2 - x} + \dots + \frac{A_{r+1}}{x_{r+1} - x} \right]. \quad (17)$$

Expanding each of the fractions into a geometric series, we get for the coefficients

$$q_{nr} = A_0 + \frac{A_1}{x_1^{n-r+1}} + \frac{A_2}{x_2^{n-r+1}} + \dots + \frac{A_{r+1}}{x_{r+1}^{n-r+1}}. \quad (18)$$

Here x_1 is the root with the smallest absolute value, and the terms with the other roots can be neglected in comparison when $n - r$ is large. For practical purposes this means all interesting cases, since $q_{nr} \approx 0$ for small $n - r$ already when $n = 8$, as Table 1 below shows. Thus we get

$$q_{nr} \approx A_0 + \frac{A_1}{x_1^{n-r+1}}. \quad (19)$$

TABLE 1.

r	x_1	$q_{8,r}$	$q_{16,r}$	$q_{32,r}$	$q_{64,r}$	$1 - e^{-2^{6-r-2}}$
0	1.333333	0.925	0.993	1.000	1.000	1.000
1	1.171573	0.710	0.921	0.994	1.000	1.000
2	1.078379	0.410	0.678	0.904	0.991	0.982
3	1.035999	0.191	0.391	0.654	0.888	0.865
4	1.016950	0.081	0.196	0.386	0.641	0.632
5	1.008197	0.032	0.093	0.204	0.387	0.393
6	1.004016	0.012	0.043	0.103	0.211	0.221
7	1.001984	0.004	0.020	0.050	0.109	0.118
8	1.000985	0.001	0.009	0.024	0.055	0.061
9	1.000491		0.004	0.012	0.027	0.031
10	1.000245		0.002	0.006	0.013	0.016
11	1.000122		0.001	0.003	0.007	0.008
12	1.000061		0.000	0.001	0.003	0.004
13	1.000031		0.000	0.000	0.002	0.002
14	1.000015		0.000	0.000	0.001	0.001
$E(M_n) = \sum_{r=0}^n q_{nr}$		2.366	3.351	4.341	5.335	5.334

We can easily evaluate the two coefficients here using Equations (8), (11) and (17), and obtain

$$q_{nr} \approx 1 - \frac{r + 1 - 2^{r+2}}{(r + 1)x_1^r - 2^{r+2}} \cdot \frac{1}{x_1^{n-r+1}} \approx 1 - \frac{1}{x_1^{n-r+1}}. \tag{20}$$

Now $E(M_n) = \sum_{r=0}^n q_{nr}$, and so this expected value can be obtained from the above formula. But, unfortunately, x_1 depends on r in a fairly complicated manner, and this precludes evaluation of the sum in closed form. Thus we present a numerical evaluation in Table 1 for $n = 8, 16, 32$ and 64 . In the last column an approximation to q_{nr} is given for $n = 64$, which we are now going to explain.

Since, with our earlier notation, $x_1 = 1 + z$, we have

$$\begin{aligned} q_{nr} &\approx 1 - \frac{1}{(1 + z)^{n-r+1}} = 1 - \left(1 + \frac{(n - r + 1)z}{n - r + 1}\right)^{-(n-r+1)} \\ &\approx 1 - e^{-(n-r+1)z} \approx 1 - \exp\left(-\frac{n - r + 1}{2^{r+2} - r - 1}\right). \end{aligned} \tag{21}$$

If we set $n = 2^k$ and neglect the linear terms, then we get

$$q_{nr} \approx 1 - e^{-2^{k-r-2}}. \tag{22}$$

Comparison of the last two columns of the table shows the amazing accuracy of this approximation. Furthermore, a Monte Carlo computer simulation for $n = 16$ produced 4.34 for the average number of iterations, in close agreement with the $1 + 3.35$ expected from the table.

The above exponential approximation to q_{nr} leads to the $\log_2 n$ estimate for $E(M_n) = \sum_{r=0}^n q_{nr}$ as follows: When $k - r \geq 5$, then $q_{nr} \approx 1$ and so, increasing k by 1 we just add an extra term of 1 to the sum. Thus

$$E(M_{2n}) \approx 1 + E(M_n) \tag{23}$$

for any n sufficiently large, and so

$$E(M_n) \approx \log_2 n + c. \quad (24)$$

From the table we find that $c \approx -0.65$, and so the expected number of iterations, being one more than this, is approximately $\log_2 n + 0.35$.

REFERENCES

1. C. C. McGeoch, Parallel Addition, *American Mathematical Monthly* 100 (1993), 867–871.
2. A. W. Burks, H. H. Goldstine, John von Neumann, Preliminary Discussion of the Logical Design of an Electronic Computing Instrument, Inst. for Advanced Study Report (1946). Reprinted in *John von Neumann Collected Works* Vol. 5, (1961).
3. N. R. Scott, *Computer Number Systems & Arithmetic*, Prentice-Hall 1985.
4. W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 1, Third Edition, Wiley, 1968.

Department of Mathematics and Computer Science
University of Massachusetts at Boston
Boston, MA 02125
gs@cs.umb.edu

How can we expect teachers to plan curriculum, choose texts, explain mathematics to their colleagues and to parents or to give a sound picture of it to their students, unless they have some knowledge and experience of the history and philosophy of mathematics and its role in contemporary culture?

—K. O. May

NOTES

Edited by: John Duncan

Fibonacci-like Sequences and Greatest Common Divisors

H. R. Morton

It is a curious feature of the Fibonacci sequence $\{f_n\}$ that the greatest common divisor (f_m, f_n) of two terms in the sequence is itself the k -th term in the sequence, with $k = (m, n)$. This result and its extension to sequences satisfying the recurrence relation

$$f_{n+1} = af_n + bf_{n-1},$$

starting with $f_0 = 0$, when a and b are any coprime integers, is proved by Lucas [L1], [L2]. The traditional proof, which is nicely presented in Hardy and Wright [HW 148–9], uses relations between the sequence $\{f_n\}$ and an auxiliary sequence, describing both sequences in terms of the roots of the quadratic $t^2 - at - b$. The purpose of this article is to present a proof which uses only simple congruence features of the sequence $\{f_n\}$. The result is stated below as Theorem A. It is deduced readily from Theorem B, which shows that the terms f_N in the sequence which are divisible by any fixed d are regularly spaced.

Theorem A. *Let $\{f_n\}$ be the sequence of integers determined by the initial conditions $f_0 = 0$, $f_1 = 1$ and the recurrence relation*

$$f_{n+1} = af_n + bf_{n-1},$$

where a and b are any two coprime integers. Then $(f_m, f_n) = \pm f_{(m,n)}$.

Remark. The choice of $f_1 = 1$ is not important; any other choice will just result in a multiple of the same sequence.

Theorem B. *Let $\{f_n\}$ be the sequence of integers defined in Theorem A. Let d be a positive integer, and let S be the set of integers N for which f_N is divisible by d . Then S consists of all multiples of some integer k , depending on d and the sequence.*

In what follows we shall use standard congruence notation and algebra; thus $f_n \equiv 0 \pmod{d}$ means that f_n is divisible by d . The only property of gcd which is needed is that every common divisor of two numbers also divides their gcd. In the case of general coprime coefficients a and b we need the result that if d divides bc and is coprime to b then d divides c , or equivalently, in the context of congruences, that any number coprime to d has an inverse mod d . In the case $b = \pm 1$ the proofs use more elementary arguments, involving only addition and multiplication mod d .

Deduction of Theorem A from Theorem B: Let d be a positive integer and let $f_m \equiv 0 \pmod{d}$ and $f_n \equiv 0 \pmod{d}$.

Consider the set S of integers N for which $f_N \equiv 0 \pmod{d}$. By Theorem B this set consists of all multiples of some k . Now $m, n \in S$, by hypothesis. Thus m and n are each divisible by k and hence also their gcd, (m, n) , is divisible by k . The integer (m, n) thus belongs to S , which in turn means that $f_{(m, n)} \equiv 0 \pmod{d}$.

Now choose $d = (f_m, f_n)$. Then f_m and f_n are both divisible by d . The argument above shows that $f_{(m, n)}$ is also divisible by $d = (f_m, f_n)$.

Conversely, choose $d = f_{(m, n)}$ and again consider the set S of integers N for which $f_N \equiv 0 \pmod{d}$. Then S consists, by Theorem B, of all multiples of some k . Clearly $(m, n) \in S$, since $f_{(m, n)}$ is divisible by d , and hence (m, n) is a multiple of k . Now m and n are multiples of (m, n) , and hence are also multiples of k . So $m, n \in S$ and thus f_m and f_n are both divisible by d . It follows at once that their gcd, (f_m, f_n) , is divisible by $d = f_{(m, n)}$.

We have already established that $f_{(m, n)}$ is divisible by (f_m, f_n) . Thus $f_{(m, n)} = \pm(f_m, f_n)$, as claimed.

It remains to establish Theorem B. This is most simply done in the case $b = \pm 1$, when a can be any integer, by extending the sequence to include the terms f_n for negative integers n also. The proof follows from two simple propositions; modifications of these needed to prove the general case are then given. Finally an alternative proof of Theorem B is indicated, along lines suggested by the referee.

Proposition 1. *Let $\{f_n\}$ be a sequence of integers satisfying the recurrence relation $f_{n+1} = af_n + bf_{n-1}$, where a and b are integers. Suppose that $f_n \equiv 0 \pmod{d}$. Then for every $k \leq n$ the terms $f_{n \pm k}$ are related by*

$$f_{n+k} + (-b)^k f_{n-k} \equiv 0 \pmod{d}.$$

Proof: By induction on k . It is clearly true for $k = 0, 1$. Now

$$\begin{aligned} f_{n+k+1} + (-b)^{k+1} f_{n-k-1} &= af_{n+k} + bf_{n+k-1} + (-b)^k af_{n-k} + b(-b)^{k-1} f_{n-k+1} \\ &\equiv 0 \pmod{d}, \end{aligned}$$

by the induction hypothesis. ■

In general, Proposition 1 shows that $f_{n+k} \equiv \pm b^k f_{n-k} \pmod{d}$ with $n \geq k$, assuming that $f_n \equiv 0 \pmod{d}$.

Suppose now that $b = \pm 1$. The relation can be read in the opposite direction as $f_{n-1} = -abf_n + bf_{n+1}$, since $b^{-1} = b$. Integers f_n satisfying the recurrence relation may then be defined for all negative integers n also. Proposition 1 holds for all k in this case, showing that $f_{n+k} \equiv \pm f_{n-k} \pmod{d}$ for all k , where $f_n \equiv 0 \pmod{d}$. Then $f_{n-k} \equiv 0 \pmod{d}$ if and only if $f_{n+k} \equiv 0 \pmod{d}$.

The set S of all integers N (positive and negative) for which $f_N \equiv 0 \pmod{d}$ is thus invariant under ‘reflection’ in any of its elements $n \in S$, where reflection in n interchanges the integers $n \pm k$.

Theorem B now follows from the geometrically obvious Proposition 2.

Proposition 2. *Any set S of integers which contains 0 and is invariant under reflection in each element of S consists of all multiples of some fixed integer k .*

Proof: Either $S = \{0\}$ or we can take $k > 0$ as the least distance between any two elements of S , which we can write as n and $n + k$. Symmetry of S under reflection in $n + k$ shows that $n + 2k \in S$. By induction on r , symmetry about $n + (r - 1)k$ shows that $n + rk \in S$ for all positive integers r . Symmetry about n extends this to show that $n + rk \in S$ for all integers r . Because k is the least distance between any two integers in S there are no further elements of S . Given that $0 \in S$ we can then write $0 = n + rk$ for some r , so that n is a multiple of k , and hence S consists of the multiples of k . ■

In the general case of coprime a and b Proposition 2 holds, when restricted to positive integers n only. In this case the reflection invariance for the set S should be taken as saying that if $n \in S$ and $n \geq k$ then $n + k \in S$ if and only if $n - k \in S$. Proposition 1 shows that $f_{n+k} \equiv \pm b^k f_{n-k} \pmod{d}$ with $n \geq k$ when $f_n \equiv 0 \pmod{d}$. Hence the set of S integers $N \geq 0$ with $f_N \equiv 0 \pmod{d}$ does have the modified reflection invariance, *provided that b and d are coprime*. Theorem B then follows in the case that d is coprime to b .

In the remaining cases, when b and d have a common factor, $c > 1$ say, the recurrence relation gives $f_{n+1} \equiv af_n \pmod{c}$, and hence $f_n \equiv a^{n-1} \pmod{c}$. Now a and b are coprime, and hence a and c are coprime, so f_n is never divisible by c for any $n > 0$. The terms f_n with $n > 0$ are then never divisible by d ; in these cases the set S consists only of 0, and again satisfies Theorem B, taking $k = 0$.

Sketch of an alternative proof of Theorem B: Observe that if $f_n \equiv 0 \pmod{d}$ then the sequence $f_n, f_{n+1}, \dots, f_{n+k}, \dots$ is a multiple of the sequence $f_0, f_1, \dots, f_k, \dots \pmod{d}$. Explicitly, an easy induction on k , using the recurrence relation, shows that $f_{n+k} \equiv f_{n+1} f_k \pmod{d}$. After another induction to prove that f_n and f_{n+1} are coprime, and hence that f_{n+1} is coprime to d , it follows that when $n \in S$ then $k \in S$ if and only if $n + k \in S$. The set S thus has the property that if $m, n \in S$ with $m \geq n$ then $m \pm n \in S$. Theorem B follows readily.

Remarks. It is interesting to look explicitly at the sequences given by small choices of a and b , besides the Fibonacci sequence with $a = b = 1$, and the integers, with $a = 2, b = -1$.

It is shown above that the terms f_n with $n > 0$ are never divisible by any prime factor of b . On the other hand Lucas showed that each prime p which is coprime to b divides some term f_n in the sequence, with $n > 0$, and hence divides infinitely many terms.

Values of n for which f_n is divisible by p can be found as follows, although these are not always the smallest possible. Set $\Delta = a^2 + 4b$ and let p be any prime not dividing Δ or b . If Δ is a square mod p then f_{p-1} is divisible by p , while if Δ is not a square mod p then f_{p+1} is divisible by p . If p divides Δ then f_p is divisible by p . Explicit details of this and other divisibility properties of Lucas are reported in [D] and [HW].

ACKNOWLEDGMENTS. This proof was developed in 1993 as a result of conversations with Rob Baston, with whom I was sharing the teaching of an elementary course involving congruences and

divisibility properties of integers. I am grateful to him and to Kit Nair and Alastair King for provoking me to complete this proof as a means of avoiding the more complicated induction proofs. I must thank the referee for suggestions which allowed me to extend my original presentation with $b = \pm 1$ to the general case, and for the outline of the alternative proof of Theorem B.

REFERENCES

- [D] L. E. Dickson, History of the Theory of Numbers, vol. 1, 1919, (Chelsea reprint, New York 1971).
- [L1] E. Lucas, Sur les rapports qui existent entre la théorie des nombres et le calcul intégral. Comptes Rendus, Paris 82 (1876), 1303–5.
- [L2] E. Lucas, Sur la théorie des nombres premiers, Atti R. Accad. Sc. Torino (Math), 11 (1875–6), 928–937.
- [HW] G. H. Hardy and E. M. Wright, Introduction to the Theory of Numbers. OUP, 1938.

Department of Pure Mathematics
University of Liverpool
P.O. Box 147
Liverpool, L69 3BX
UNITED KINGDOM
h.r.morton@liv.ac.uk

Generating Symmetric Groups

I. M. Isaacs and Thilo Zieschang

It is well known that the symmetric group S_n on the symbols $\{1, 2, 3, \dots, n\}$ can be generated by two carefully chosen permutations. It is easy to check, for example, that the cycles $x = (1, 2)$ and $y = (1, 2, 3, \dots, n)$ will do the job. We prove in this note that except when $n = 4$, care is needed in the choice of only one of the two generators.

Theorem A. *Assume that $n \neq 4$ and let $x \in S_n$ be an arbitrary nonidentity element. Then there exists an element $y \in S_n$ such that $S_n = \langle x, y \rangle$.*

We mention that when $n = 4$, the conclusion of Theorem A really fails. If $x = (1, 2)(3, 4)$, then x lies in the normal Klein subgroup K of S_4 of order 4. Since the factor group S_4/K is noncyclic, there can be no element $y \in S_4$ such that $\langle x, y \rangle$ is the whole group.

To prove Theorem A, we need a way to recognize when a subgroup $G \subseteq S_n$ is actually the whole group. A well-known (and nearly trivial) result of this type is that if G contains all transpositions (that is 2-cycles) of S_n , then $G = S_n$. It is almost as easy to see also that if G contains all 3-cycles of S_n , then it contains the alternating group A_n , and so either $G = A_n$ or $G = S_n$. In this case, if we can find some odd permutation in G , it follows that G is the whole group S_n .

To use the results of the previous paragraph, it may seem necessary to undergo the tedium of checking that the subgroup G contains every transposition or every 3-cycle. There is a marvelous short-cut, however, discovered around 1870 by

C. Jordan, that enables one to get away with establishing the existence of just one transposition or one 3-cycle in G . Obviously, this could not possibly work for a completely arbitrary subgroup $G \subseteq S_n$, and there is another hypothesis needed for Jordan's theorem.

Write $\Omega = \{1, 2, 3, \dots, n\}$. If $x \in S_n$ and $\Delta \subseteq \Omega$, we write Δx to denote the image of Δ under the map x . (The subset $\Delta x \subseteq \Omega$ is called the **translate** of Δ under the permutation x .) Now fix a subgroup $G \subseteq S_n$. A nonempty subset $\Delta \subseteq \Omega$ is said to be a **block** for G if for each element $x \in G$, the translate Δx is either disjoint from or equal to Δ . Clearly, each singleton subset of Ω is a block and so too is the whole set Ω , but these are certainly not very interesting and they are referred to as **trivial** blocks. The situation in which Jordan's theorem applies is where the group G is **primitive**, which means that the *only* blocks for G are the trivial ones.

It is instructive to play a little with the definitions of blocks and primitive groups. Fix a subgroup $G \subseteq S_n$ and observe that if the set Ω can be decomposed into pairwise disjoint parts that are permuted by the translations via elements of G , then each part is a block. (We call such a decomposition of Ω a **G -invariant partition**.) Conversely, every block for G must be one of the parts of some **G -invariant partition**. To see this, let Δ be an arbitrary block and observe that the translates of Δ via elements of G must also be blocks. The distinct translates of Δ are pairwise disjoint, therefore, and if their union is the whole set Ω , we have a G -invariant partition. Otherwise, we can get a G -invariant partition by creating one additional part consisting of all the left-over points.

We can always decompose Ω into its orbits under the action of G and we observe that this is trivially a G -invariant partition. Orbits are thus blocks and it follows that if G is primitive, then either all orbits are singleton sets and G is the trivial group, or else the whole set Ω is an orbit and G is transitive. It is easy to see for $n > 2$ that the trivial group is not primitive and it follows (for $n \neq 2$) that primitive groups are always transitive.

To see a natural example of a transitive group that is not primitive, imagine marking the faces of a cube with the numbers 1 through 6 and let $G \subseteq S_6$ be the group of permutations induced by rotations of the cube. (Note that G is transitive and $|G| = 24$.) For definiteness, suppose that the cube is numbered as is standard for dice, so that on each pair of opposite faces, the numbers total 7. Since every rotation of the cube carries a pair of opposite faces to a pair of opposite faces, we see that the three sets $\{1, 6\}$, $\{2, 5\}$ and $\{3, 4\}$ form a G -invariant partition of $\Omega = \{1, 2, 3, 4, 5, 6\}$, and hence each of them is a nontrivial block for G , which is therefore imprimitive.

In general, if Δ is a block of a transitive subgroup $G \subseteq S_n$, then the G -translates of Δ cover Ω , and hence they form a G -invariant partition in which all parts have equal size. It follows in this case that $|\Delta|$ must divide n . Also, if Δ is nontrivial, then so are its translates, and hence if G is transitive but imprimitive, it follows that every element of Ω lies in a nontrivial block. We state this observation formally for future reference.

Lemma. *Suppose $G \subseteq S_n$ is transitive and let $a \in \Omega$. Then G is primitive if the only blocks containing a are $\{a\}$ and Ω .* ■

One of the goals of this paper is to provide a direct and elementary proof of Jordan's theorem, which we can now state.

Theorem (Jordan). *Suppose that G is a primitive subgroup of S_n .*

- (a) *If G contains a transposition, then $G = S_n$.*
- (b) *If G contains a 3-cycle, then either $G = S_n$ or $G = A_n$.*

Proof: We prove part (a) first. Build an undirected graph \mathcal{G} with vertex set $\Omega = \{1, 2, 3, \dots, n\}$ by joining distinct vertices a and b if the transposition (a, b) happens to lie in the group G . The connected components of \mathcal{G} partition the vertex set Ω and we claim that these components are blocks for G . It suffices to show that the components form a G -invariant partition, and so we must prove that they are permuted by the elements of G . Since it is clear that the components are permuted by graph automorphisms, we want to show that each element of G actually is an automorphism of \mathcal{G} .

If vertices a and b are joined in \mathcal{G} , we must show for each element $g \in G$ that vertices $(a)g$ and $(b)g$ are also joined. If a and b are joined, however, then the transposition $t = (a, b)$ lies in G and hence $t^g = g^{-1}tg$ is also an element of G . Since t^g is the transposition $((a)g, (b)g)$, however, we deduce that $(a)g$ and $(b)g$ actually are joined in \mathcal{G} , as required.

We now know that the connected components of \mathcal{G} are blocks for G . By assumption G is primitive, however, and this tells us that either each component is a singleton and the graph is totally disconnected, or else the whole set Ω is one component and the graph is connected. Since we are given that G contains a transposition, we know that \mathcal{G} contains an edge and it is not totally disconnected. It follows that \mathcal{G} is a connected graph.

To prove that G is the full symmetric group, it suffices to show that it contains an arbitrary transposition (a, b) . Seeking a contradiction, we assume that vertices a and b are not directly joined in \mathcal{G} . We know that there is some path leading from a to b in the graph and we suppose that a, m and n are three consecutive vertices in some shortest path from a to b . (Note the possibility that $n = b$.) Since transpositions (a, m) and (m, n) are in G , it follows that $(a, n) = (m, n)(a, m)(m, n)$ is also an element of G , and thus a is joined directly to n in \mathcal{G} . This is a contradiction since it follows that we can delete m from a shortest path from a to b to obtain a still shorter path.

The proof of (b) is similar, but a little more complicated. Again we construct an undirected graph \mathcal{G} with vertex set Ω , but this time, we join vertices a and b if G contains some 3-cycle moving both a and b . (In other words, a and b are joined iff G contains both the 3-cycle (a, b, u) and its inverse (b, a, u) for some point $u \in \Omega$.) Here too, the permutations $g \in G$ are graph automorphisms since if $t = (a, b, u)$ lies in G , then $t^g = ((a)g, (b)g, (u)g)$ also lies in G . As in the proof of part (a), the hypotheses on G enable us to deduce that the graph \mathcal{G} is connected.

Continuing to parallel the proof of part (a), we show next that \mathcal{G} is a complete graph. Exactly as before, it suffices to show that if a, m and n are three distinct vertices such that a is joined to m and m is joined to n , then a and n are directly joined. We know that G contains a 3-cycle g that moves a and m and a 3-cycle h that moves m and n , and our task is to produce a 3-cycle in G that moves a and n . We are done unless $(n)g = n$ and we can assume that $g = (m, a, u)$ so that $(m)g = a$. Now h is a 3-cycle moving m and n and it follows that its conjugate h^g is a 3-cycle moving $(m)g = a$ and $(n)g = n$, as desired.

To show that G is either A_n or S_n , it suffices to show that G contains an arbitrary 3-cycle (a, b, c) . Since the graph \mathcal{G} is known to be complete, vertices a

and b are joined and thus G contains the 3-cycle $t = (a, b, u)$ for some element $u \in \Omega$. Similarly, G contains $s = (b, c, v)$ and we can certainly assume that $u \neq c$ and that $v \neq a$. If $u = v$, then $st = (b, c, u)(a, b, u) = (a, b, c)$ and this lies in G , as required. If $u \neq v$, on the other hand, we compute that

$$t^{-1}s^{-1}tst = (u, b, a)(v, c, b)(a, b, u)(b, c, v)(a, b, u) = (a, b, c)$$

and again, $(a, b, c) \in G$. ■

The following result has appeared in various places as a problem, apparently with the intention that it should be done by ‘brute force’. In fact, it provides a good demonstration of the power of Jordan’s theorem, and that is why we present it here.

Theorem B. *In the symmetric group S_n , write $x = (1, 2, 3, \dots, n)$ and let $y = (1, 2, 3, \dots, m)$ for some integer m such that $1 < m < n$. Then $\langle x, y \rangle$ is the whole symmetric group unless both n and m are odd, in which case $\langle x, y \rangle$ is the alternating group A_n .*

Proof: We show first that $G = \langle x, y \rangle$ is primitive. Certainly, G is transitive, and so by the lemma, it suffices to show that a block Δ containing 1 and at least one other number $a \in \Omega$ must be the whole set Ω . If $a > m$, then $(a)y = a$, and so $a \in \Delta \cap \Delta y$. But Δ is block, and hence $\Delta y = \Delta$. Since $1 \in \Delta$, we see that $2 \in \Delta$, and thus $2 \in \Delta \cap \Delta x$. We conclude that $\Delta x = \Delta$ and thus Δ must be the whole set Ω , as desired.

If, on the other hand, $a \leq m$, then since $a > 1$, we see that $a = (a)x^{-1}y$. Thus $a \in \Delta \cap \Delta x^{-1}y$ and we conclude that $\Delta x^{-1}y = \Delta$. Thus $n = (1)x^{-1}y$ lies in Δ and we are in the case of the previous paragraph.

Since we now know that G is primitive, we will be able to apply Jordan’s theorem if we can find a 3-cycle in G . Observe that

$$\begin{aligned} (a)xy &= a + 2 = (a)yx & \text{if } 1 \leq a \leq m - 2 \text{ and} \\ (a)xy &= a + 1 = (a)yx & \text{if } m + 1 \leq a \leq n - 1, \end{aligned}$$

and so xy and yx agree on all numbers in Ω except possibly $m - 1$, m and n . It follows that $xyx^{-1}y^{-1}$ fixes all but these three numbers and we compute that

$$\begin{aligned} (m - 1)xyx^{-1}y^{-1} &= n, \\ (m)xyx^{-1}y^{-1} &= m - 1 \text{ and} \\ (n)xyx^{-1}y^{-1} &= m. \end{aligned}$$

It follows that $xyx^{-1}y^{-1} = (m - 1, n, m)$ and G does contain a 3-cycle.

By Jordan’s theorem, G is either A_n or S_n and our remaining task is to determine which group we actually have. If n is even, then x is an odd permutation and if m is even, then y is an odd permutation, and so in these cases $G \neq A_n$ and we conclude $G = S_n$. If m and n are both odd, however, then x and y are even permutations, which lie in A_n . It follows that $G \subseteq A_n$ and hence $G = A_n$ in this case. ■

Proof of Theorem A. The result is clear when $n < 4$, and so we can assume that $n > 4$ and we consider first the case where n is odd. By renaming the symbols being permuted, we can suppose that x moves 1 but that $(1)x \neq 2$. Let $y =$

$(1, 2)(3, 4, \dots, n)$, the product of a transposition and a cycle having odd length $n - 2$ and write $G = \langle x, y \rangle$. Since $n - 2$ is odd, the element y^{n-2} is a transposition in G and it suffices by Jordan's theorem to show that G is primitive.

Since x carries 1 to something other than 2, we see that G is transitive. Let $\Delta < \Omega$ be a block containing 1, so that by the lemma, it suffices to show that $\Delta = \{1\}$. Note that $|\Delta|$ is a proper divisor of n and in particular, it is odd and at most $n/2$. Since y^2 fixes 1, we see that $1 \in \Delta \cap \Delta y^2$, and thus $\Delta = \Delta y^2$. But $\{3, 4, \dots, n\}$ is an orbit for $\langle y^2 \rangle$ (since $n - 2$ is odd), and thus if any one of the numbers a with $3 \leq a \leq n$ lies in Δ , they all do, and $|\Delta| \geq n - 1 > n/2$, a contradiction. Also $2 \notin \Delta$ since otherwise $|\Delta| = 2$, and this is a contradiction too. Thus $\Delta = \{1\}$ and we are now done in the case where n is odd.

Now, assume that n is even. If x is a transposition, we can suppose that $x = (1, 2)$ and we take $y = (1, 2, 3, \dots, n)$ so that $\langle x, y \rangle$ is the whole symmetric group. If x is a 3-cycle, we can suppose that $x = (1, 2, 3)$ and again we take $y = (1, 2, 3, \dots, n)$. In this case too, $\langle x, y \rangle$ is the whole symmetric group by Theorem B, since n is even.

We can now assume that x moves at least four points. By renaming symbols if necessary, we can suppose that $(3)x = 4$. There are at least two numbers other than 3 and 4 moved by x and at least one of these, say 1, is not carried to 3 and we can assume $(1)x = 2$. Now let $y = (2, 3)(4, 5, \dots, n)$, the product of a transposition and a cycle of odd length $n - 3$, and let $G = \langle x, y \rangle$. Since $n - 3$ is odd, y^{n-3} is a transposition in G and by Jordan's theorem, it suffices to show that G is primitive.

Since x carries 1 to 2 and 3 to 4, we see that G is transitive on Ω . The lemma thus applies, and so as before, if we suppose that $\Delta < \Omega$ is a block for G containing 1, it suffices to show that $\Delta = \{1\}$. Now y fixes 1, and so $1 \in \Delta \cap \Delta y$ and we conclude that $\Delta y = \Delta$. It follows that if any one of the numbers a with $4 \leq a \leq n$ lies in Δ , they all do. In this situation, $|\Delta| \geq n - 2 > n/2$, where the strict inequality holds because $n > 4$. This is a contradiction since $|\Delta|$ is a proper divisor of n , and we conclude that $\Delta \subseteq \{1, 2, 3\}$.

Recall that $(1)x = 2$ and $(3)x = 4$ and hence $\{1, 2, 3\}x$ is neither equal to nor disjoint from $\{1, 2, 3\}$. Thus $\{1, 2, 3\}$ is *not* a block for G , and so Δ must be a proper subset of this set. Because $\Delta y = \Delta$, however, we see that if either 2 or 3 lies in Δ , they both do, and this is a contradiction. We conclude that $\Delta = \{1\}$, as required. ■

A result similar to Theorem A is known to be valid for the alternating group A_n for all values of n . Although it seems likely that a proof of this result along the lines of our proof of Theorem A might exist, there are technical difficulties in some cases, and we have not actually found such a proof.

Finally, we remark that Jordan proved much more than the result we credited him with here. He showed that if G is a primitive subgroup of S_n and H is a nontrivial subgroup of G that fixes m points and is primitive in its action on the remaining $n - m$ points, then G is $(m + 1)$ -fold transitive. (This means that given two arbitrary ordered $(m + 1)$ -tuples of distinct points of Ω , there exists an element of G that carries one to the other.) The result of Jordan that we stated follows easily from this by taking H to be the subgroup generated by the given transposition or 3-cycle. Much more can be obtained, however. For example, suppose that instead of a transposition or a 3-cycle, we know that G contains a p -cycle for some arbitrary prime number p . It is not too hard to show from Jordan's result that if $p \leq n - 3$, then G must be either A_n or S_n . We refer the reader to Wielandt's book [1] for more information on all of this.

1. H. Wielandt, *Finite Permutation Groups*, Academic Press, New York, 1964.

*Department of Mathematics
University of Wisconsin
Madison, WI 53706-1313
isaacs@math.wisc.edu*

*Department of Computer Science
University of Saarland
Geb. 36, Postfach 1150
66041 Saarbruecken
zie@cs.uni-sb.de*

On the Arithmetic–Geometric Mean Inequality

Lutz G. Lucht

Beckenbach and Bellman [1] contains many beautiful proofs of the well-known inequality between the weighted arithmetic and geometric means of n positive real numbers. In this note another short proof is given which is based on the common log properties: (i) the log curve is concave, (ii) the log function is a homomorphism of \mathbb{R}_+ onto \mathbb{R} .

Let t be a positive real number. Then

$$t - 1 > \log t \tag{1}$$

except for $t = 1$ when equality in (1) obviously holds. This comes, for example, from the mean-value theorem in analysis by considering the logarithmic function on the interval with endpoints 1, t .

Suppose that the real numbers ξ, x_1, \dots, x_n and the weights $\lambda_1, \dots, \lambda_n$ are positive, with $\lambda_1 + \dots + \lambda_n = 1$. From (1), applied to $t = x_\nu/\xi$, we obtain after multiplication with $\lambda_\nu \xi$

$$\lambda_\nu x_\nu \geq \lambda_\nu \xi + \xi \log \frac{x_\nu^{\lambda_\nu}}{\xi^{\lambda_\nu}} \quad (\nu = 1, \dots, n).$$

Addition gives

$$\lambda_1 x_1 + \dots + \lambda_n x_n \geq \xi + \xi \log \frac{x_1^{\lambda_1} \dots x_n^{\lambda_n}}{\xi}.$$

Now choose

$$\xi = x_1^{\lambda_1} \dots x_n^{\lambda_n},$$

and the arithmetic-geometric mean inequality

$$\lambda_1 x_1 + \dots + \lambda_n x_n \geq x_1^{\lambda_1} \dots x_n^{\lambda_n} \tag{2}$$

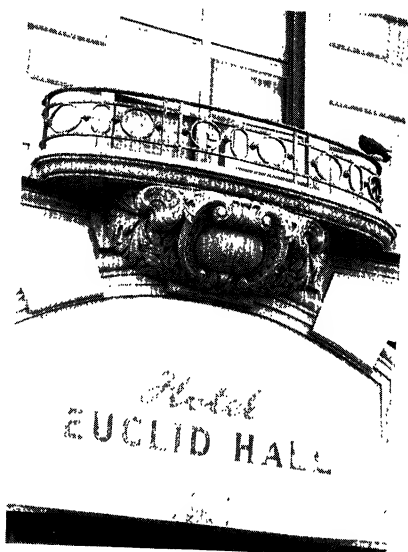
follows. The above remark concerning equality in (1) shows that the inequality (2) is strict unless $x_1 = \dots = x_n$.

We notice that instead of characterizing (i) by the property *the chords are under the curve* the proof uses *the curve is under the tangent lines*. Though by (1) only one tangent line seems to be involved, via (ii) all tangent lines are actually used.

REFERENCES

1. E. F. Beckenbach and R. Bellman, *Inequalities*, Springer-Verlag, Berlin Heidelberg New York Tokyo, 1983.

*Institut für Mathematik
Technische Universität Clausthal
38678 Clausthal-Zellerfeld, Germany
lucht@math.tu-clausthal.de*



Hotel Euclid Hall
2345 Broadway
New York City



Hotel Newton
2528 Broadway
New York City

*Submitted by Jonathan D. Sondow
209 W. 97 St., Apt 6F
New York, NY 10025*

UNSOLVED PROBLEMS

Edited by: Richard Guy & Richard Nowakowski

In this department the MONTHLY presents easily stated unsolved problems dealing with notions ordinarily encountered in undergraduate mathematics. Each problem should be accompanied by relevant references (if any are known to the author) and by a brief description of known partial or related results. Typescripts should be sent to Richard Guy, Department of Mathematics & Statistics, The University of Calgary, Alberta, Canada T2N 1N4.

Three Open Problems in Functional Equations

P. K. Sahoo

In this note we seek solution of three problems connected with the characterizations of sum form information measures on open domain. The first problem is the following: Find all functions $f: (0, 1) \rightarrow \mathfrak{R}$ (the set of reals) satisfying the functional equation

$$f(xy) + f(x(1-y)) + f(y(1-x)) + f((1-x)(1-y)) = 0 \quad (1)$$

for all $x, y \in (0, 1)$. This problem was stated as an open problem in [6]. If f is assumed to be measurable, then Daroczy and Jarai [5] have shown that $f(x) = 4ax - a$, where a is an arbitrary constant. Recently, Maksa [13] has posed the following problem at the Thirtieth International Symposium on Functional Equations: Find all functions $f: [0, 1] \rightarrow \mathfrak{R}$ satisfying the functional equation

$$(1-x-y)f(xy) = xf(y(1-x)) + yf(x(1-y)) \quad (2)$$

for all $x, y \in [0, 1]$. One can easily show that if f is a solution of (2), then f is skew symmetric about $\frac{1}{2}$, that is $f(x) = -f(1-x)$, and $f(0) = 0$. Further, it is easy to note that Maksa's equation (2) implies equation (1). To see this replace x by $1-x$ in (2) and add the resulting equation to (2) to obtain

$$y[f(xy) + f(x(1-y)) + f(y(1-x)) + f((1-x)(1-y))] = 0$$

for all $x, y \in (0, 1]$. Since $f(0) = 0$, the above equation yields (1) for all $x, y \in [0, 1]$. Thus, the general solution of (1) will provide the general solution of (2). Utilizing the solution given by Daroczy and Jarai [5] of the equation (1), it is easy to show that if f is measurable or almost open, then all solutions of (2) are of the form $f(x) = 0$.

The second problem is the following: Find all functions $f: (0, 1) \rightarrow \mathfrak{R}$ satisfying the functional equation

$$f(xy) + f((1-x)(1-y)) = f(x(1-y)) + f(y(1-x)) \quad (3)$$

for all $x, y \in (0, 1)$. Daroczy and Jarai [5] have also found the measurable solution

of this functional equation. They have shown that the measurable solution of (3) is of the form $f(x) = ax^2 - ax + b \log x + c$, where a , b and c are arbitrary constants. Equation (3) appears in [11] as a problem posed by Lajko when (3) holds for all x and y in \mathfrak{R} . Eliezer [7] has determined the differentiable solution of Lajko's problem. Eliezer proved that if f is differentiable and satisfies (3) for all x and y in \mathfrak{R} , then $f(x) = ax^2 - ax + c$, where a and c are arbitrary constants.

Finally, our last problem is the following: Find all functions $f, g, h: (0, 1) \rightarrow \mathbb{R}$ satisfying the functional equation

$$f(xy) + f(x(1-y)) + f(y(1-x)) + f((1-x)(1-y)) = g(x)h(y) \quad \forall x, y \in (0, 1). \quad (4)$$

This functional equation also arises in the characterizations of information measures (see [6] and [12]) and equation (1) is a special case of it. The measurable complex-valued solution of this equation has been obtained by Losonczi in [12]. Interested readers are referred to [1, 2, 3, 4, 8, 9, 10, 14, 15] for treatments of the general subject.

REFERENCES

1. J. Aczel, *Lectures on Functional Equations and Their Applications*. Academic Press, New York-London, 1966.
2. J. Aczel, *A Short Course on Functional Equations Based upon Recent Applications to the Social and Behavioral Sciences*. D. Reidel Publishing Co., Dordrecht-Boston-Lancaster-Tokyo, 1987.
3. J. Aczel and J. Dhombres, *Functional Equations in Several Variables*. Cambridge University Press, Cambridge, 1989.
4. E. Castillo and M. R. Ruiz-Cobo, *Functional Equations and Modelling in Science and Engineering*. Marcel Dekker, Inc., New York-Basel-Hong Kong, 1992.
5. Z. Daroczy and A. Jarai, On the measurable solution of a functional equation of the information theory. *Acta Math. Acad. Sci. Hungaricae*, 34 (1979) 105-116.
6. B. R. Ebanks, P. K. Sahoo and W. Sander, Determination of measurable sum form information measures satisfying $(2, 2)$ -additivity of degree (α, β) *Radovi Matematički*, 6 (1990) 77-96.
7. C. J. Eliezer, A solution to $f((1-x)(1-y)) + f(xy) = f(x(1-y)) + f(y(1-x))$. *Aequationes Math.*, 10 (1974) 311-312.
8. P. Kannappan and C. T. Ng, On a functional equations and measures of information I. *Acta Math. Acad. Sci. Hungaricae*, 40 (1985) 243-249.
9. M. Kuczma, *An Introduction to the Theory of Functional Equations and Inequalities. Cauchy's Equation and Jensen's Inequality*. Prace Nauk. Uniw. Śl. 489, Polish Scientific Publishers, Warsaw-Cracow-Katowice, 1985.
10. M. Kuczma, B. Choczewski and R. Ger, *Iterative Functional Equations*. Cambridge University Press, Cambridge, 1990.
11. K. Lajko, Problem, P116, *Aequationes Math.*, 10 (1974) 311.
12. L. Losonczi, Measurable solutions of a functional equation related to $(2, 2)$ -additivity entropies of degree α . *Publ. Math. Debrecen*, 42 (1993) 109-137.
13. Gy. Maksa, Problem, *Aequationes Math.*, 46 (1993) 301.
14. J. Smital, *On Functions and Functional Equations*. Adam Hilger, Bristol-Philadelphia, 1988.
15. L. Székelyhidi, *Convolution Type Functional Equations on Topological Abelian Groups*. World Scientific, Singapore-New Jersey-London-Hong Kong, 1991.

Department of Mathematics
University of Louisville
Louisville, KY 40292
pk_saho01@homer.louisville.edu

PROBLEMS AND SOLUTIONS

Edited by:
Richard T. Bumby, Fred Kochman and Douglas B. West

Proposed problems should be sent to the MONTHLY PROBLEMS address given on the inside front cover. Please include solutions and relevant references. Three copies of all items needed to evaluate the problem should be sent.

Solutions of published problems should arrive at the MONTHLY PROBLEMS address given on the inside front cover before March 31, 1996. If possible, solutions should be typed with double spacing. Two copies suffice. Several solutions may be mailed together, but they should be on separate sheets of paper. The problem number and the solver's name and mailing address should appear on each solution. A mailing label should be included if an acknowledgment is desired.

The published solution is likely to be based on a solution that is complete and correct. Additional information, such as references to other appearances of the problem or its solution, is also welcome.

An asterisk () after the number of a problem, or part of a problem, indicates that no solution is currently available.*

PROBLEMS

10473. *Proposed by Emre Alkan (student), Bosphorus University, İstanbul, Turkey.*

Prove that there are infinitely many positive integers m such that

$$\frac{1}{5 \cdot 2^m} \sum_{k=0}^m \binom{2m+1}{2k} 3^k$$

is an odd integer.

10474. *Proposed by Harry Tamvakis (student), The University of Chicago, Chicago, IL.*

Consider a triangle ABC and a point P in the interior of ABC , and let the lines AP , BP , CP meet the lines BC , CA , AB at the points D , E , F respectively. Show that $\angle EDF$ is a right angle if and only if

$$\frac{1}{|PD|} = \frac{1}{|AD|} + \frac{1}{|BD|} + \frac{1}{|CD|}.$$

10475. *Proposed by Wu Wei Chao, He Nan Normal University, Xin Xiang City, He Nan Province, China.*

For $0 < x < y < 1$ or $1 < x < y$, prove that

$$y^{x^y} / x^{y^x} > y / x > y^x / x^y.$$

10476. *Proposed by Simeon T. Stefanov, Sofia, Bulgaria.*

Let \mathbf{X} be a countable compact Hausdorff space. Prove that every continuous map $f: \mathbf{X} \rightarrow \mathbf{X}$ has a periodic point.

10477. *Proposed by B. H. Neumann, Australian National University, Canberra, Australia, and D. G. Rogers, University of Aberdeen, Aberdeen, Scotland.*

Let S be a subset of an abelian group A with the composition operator $+$ and assume that S is closed under negation. If f is a permutation of S , construct the functions If , Rf and Cf from S to A as follows. If $f(i) = j$, define

$$If(j) = i; \quad Rf(-i) = -j; \quad Cf(i) = i - j.$$

Let e be the identity function on S , and call a permutation p of S *complete* if $q = p - e$ is again a permutation of S .

(a) Show that, if p is a complete permutation of S , then so are Ip , Rp and Cp , and hence that I , R and C may be viewed as operators on the set of complete permutations of S .

(b) Show that I , R and C are involutions, and determine the largest group that they can generate.

(c) If p is a complete permutation of S , show that the function Bp defined by $Bp = p \circ Iq$ is also a complete permutation. Does the involution B defined in this way lie in the group generated by I , R and C ?

10478. *Proposed by Joan P. Hutchinson, Macalester College, Saint Paul, MN.*

Let P be a simple closed n -gon, not necessarily convex (an “art gallery”), with some pairs of vertices joined by nonintersecting interior diagonals (“walls”), and suppose that in the interior of each of these diagonals there is an arbitrarily placed, arbitrarily small opening (a “doorway”). Determine the size of the smallest set G of points (“guards”) so that for every other point q in P there is a line segment in P , disjoint from the punctured diagonals, that joins q to a point of G .

10479. *Proposed by Jeffrey C. Lagarias, AT&T Bell Laboratories, Murray Hill, NJ, and Bjorn Poonen, University of California, Berkeley, CA.*

Let p be an odd prime, and consider the polynomial

$$F_p(x) = \sum_{j=1}^{p-1} j^{\frac{p-1}{2}} x^j$$

with coefficients modulo p . Show that, as a polynomial over the finite field with p elements, $F_p(x)$ has a root at $x = 1$ of multiplicity exactly $\frac{p-1}{2}$.

NOTES

(10476) A point $x \in X$ is called a *periodic point* of a map $f: X \rightarrow X$ if $f^p(x) = x$ for some integer $p \geq 1$. (10477) Complete permutations are similar to the *graceful permutations* of $[1, \dots, n]$ of E 3455 [1991, 646; 1992, 691]. The involutions I and R are similar to the involutions π and ρ in the solution of that problem. (10478) The *Art Gallery Theorem* states that an n sided art gallery can be guarded by $\lfloor n/3 \rfloor$ guards, and this bound is best possible (see J. O'Rourke, *Art Gallery Theorems and Algorithms*, Oxford University Press, 1987). However, real art galleries have opaque interior walls. This problem adds that feature. The guards may stand in doorways and survey both adjacent rooms; lines of sight may pass along walls. (10479) While this problem considers the $F_p(x)$ as polynomials over a finite field, the coefficients are *Legendre symbols*, and could be written as ± 1 for $0 < j < p$. Such polynomials with coefficients ± 1 were introduced by M. Fekete in 1912 in a study of zeros of Dirichlet L -functions. Classically, the roots of these polynomials in the real numbers have been studied. See G. Pólya "Verschiedene Bemerkungen zur Zahlentheorie" (Collected Works, Analysis, pp. 76–85 and 488) for more information.

SOLUTIONS

A Special Value of a Quadratic Form

10258[1992, 873]. *Proposed by Hans Liebeck and Anthony Osborne, University of Keele, England.*

Let a, b , and c be positive integers which are pairwise relatively prime. Prove that if the congruences

$$A^2 \equiv -bc \pmod{a}, \quad B^2 \equiv -ca \pmod{b}, \quad C^2 \equiv -ab \pmod{c}$$

are solvable for A, B , and C , then the equation

$$ax^2 + by^2 + cz^2 = abc$$

has a solution in integers x, y , and z .

Solution by Robin J. Chapman, University of Exeter, Exeter, U. K. Define u, v, w by $u \equiv (C/a) \pmod{c}$, $v \equiv (B/a) \pmod{b}$, and $w \equiv (A/b) \pmod{a}$. Let

$$\Lambda = \left\{ (x, y, z) \in \mathbb{Z}^3 : x \equiv uy \pmod{c}, x \equiv vz \pmod{b}, y \equiv wz \pmod{a} \right\}.$$

It is clear that Λ is a sublattice of \mathbb{Z}^3 . If $(x, y, z) \in \Lambda$ and z is given, then the congruence class of y modulo a is determined, and then so are the congruence classes of x modulo b and modulo c , i.e., modulo bc . Hence Λ has index abc in \mathbb{Z}^3 . Since $w^2b \equiv -c \pmod{a}$, we also have $ax^2 + by^2 + cz^2 \equiv b(wz)^2 + (-bw^2)z^2 \equiv 0 \pmod{a}$. Similarly, $ax^2 + by^2 + cz^2$ is also divisible by b and c .

Let

$$E = \left\{ (x, y, z) \in \mathbb{R}^3 : ax^2 + by^2 + cz^2 < 2abc \right\}.$$

This is an open ellipsoidal ball of volume

$$V = \frac{4\pi(2abc)^{3/2}}{3\sqrt{abc}} = \frac{8\pi\sqrt{2}abc}{3}.$$

Since $\pi\sqrt{2}/3 > 1$, we have $V > 8abc$. Now Minkowski's Theorem guarantees a nonzero (x, y, z) in $\Lambda \cap E$. With $0 < ax^2 + by^2 + cz^2 < 2abc$ and $ax^2 + by^2 + cz^2$ divisible by abc , we have $ax^2 + by^2 + cz^2 = abc$, as desired.

Solved also by I. Kastanas, A. D. Melas (Greece), F. Schmidt, GCHQ Problem Solving Group (U. K.), and the proposers. One incorrect solution and one incomplete solution were also received.

Powers of the Symmetric Group

10267[1992, 958]. *Proposed by Lenny Jones and Mike Seyfried, Shippensburg University, Shippensburg, PA, and Stephen Schroer, Mercersburg Academy, Mercersburg, PA.*

Find all pairs of positive integers $\langle n, k \rangle$ such that the set of all k th powers of elements of the symmetric group S_n on n things is a proper subgroup of S_n .

Solution by National Security Agency Problems Group, Fort Meade, MD. The only such pairs (with k less than the exponent of S_n) are $(3, 2)$, $(3, 4)$, $(4, 2)$, $(4, 6)$, $(4, 10)$, and the pairs $(5, 2l)$ such that l is relatively prime to 60 and $2l$ is less than 60.

Suppose that G_k , the set of all k th powers in S_n , is a proper subgroup. Since G_k is invariant under conjugation, it must be a normal subgroup.

Suppose first that $n \geq 6$. In this case, the only proper normal subgroup of S_n is the alternating group A_n , so $G_k = A_n$. The k th power of the transposition (12) must lie in $G_k = A_n$, so k is even. Thus all elements of G_k are squares. The even permutation with cycle representation $(1234)(56)$ is not a square, since its square root would have order 8, but the exponent of S_6 being $3 \times 4 \times 5$ prohibits elements of order 8. Therefore, there are no solutions for $n \geq 6$.

Now suppose that $n \leq 5$. Since S_1 and S_2 have no proper subgroups, we require $n \geq 3$. Without loss of generality, we assume that k is less than the exponent of S_n . A direct search reveals that $(n, k) = (3, 2)$, $(3, 4)$, $(4, 2)$, $(4, 6)$, $(4, 10)$ are the only solutions with $n \leq 4$. The exponent of S_5 is 60, and the argument of the first paragraph implies that G_k is a proper subgroup only if $G_k = A_5$. As above, we conclude that k is even. Thus $G_k \subseteq G_2$, but also $G_2 \subseteq A_5$, so $G_k = G_2$. Since A_5 is the set of squares of elements of S_5 , $(5, 2)$ is a solution. If l is relatively prime to the exponent 60 of S_5 , then the map sending x to x^l is injective. Hence $G_{2l} = G_2 = A_5$, and the pairs described above are all solutions.

It remains to show that $G_{2l} \neq A_5$ if $\gcd(l, 60) \neq 1$. If l is even, then $G_{2l} \subseteq G_4$, but $(12)(34) \in A_5$ is not a fourth power; its fourth root would have order 8, but no element of S_5 has order greater than 6. Similarly, if 3 divides l , then $G_{2l} \subseteq G_6$, but $(123) \in A_5$ is not a sixth power. Lastly if 5 divides l , then $G_{2l} \subseteq G_{10}$, but $(12345) \in A_5$ is not a tenth power.

Solved also by R. J. Chapman (U. K.), G. Ehrlich, S. M. Gagola Jr., O. P. Lossers (The Netherlands), F. Schmidt, GCHQ Problem Solving Group (U. K.), and the proposers. One incorrect solution was received.

Splitting a Sequence of Ultrafilters

10273[1992, 958]. *Proposed by Jesús Ferrer, Universidad de Valencia, Burjasot, Spain.*

Let $\langle \mathcal{U}_n \rangle$ be a sequence of distinct ultrafilters on the set \mathbb{N} of non-negative integers.

(a) Show that there is a sequence of disjoint sets $\langle A_k \rangle$ such that each A_k is an element of some \mathcal{U}_n .

(b) Show that there is $M \subset \mathbb{N}$ such that

$$\{n \in \mathbb{N} : M \in \mathcal{U}_n\} \quad \text{and} \quad \{n \in \mathbb{N} : M \notin \mathcal{U}_n\}$$

are both infinite.

Solution by Timothy J. LaBerge, Union College, Schenectady, NY. Let $\langle \mathcal{U}_n \rangle$ be a sequence of distinct ultrafilters on \mathbb{N} . For a subset A of \mathbb{N} , we let A^c denote the relative complement $\mathbb{N} \setminus A$ of A in \mathbb{N} . We begin by recursively constructing a subsequence $\langle \mathcal{U}_{n_k} \rangle$ of $\langle \mathcal{U}_n \rangle$ and a sequence $\langle A_n \rangle$ of pairwise disjoint infinite subsets of \mathbb{N} satisfying

1. $A_k \in \mathcal{U}_{n_k}$ and

2. $\{n \in \mathbb{N} : B_k \in \mathcal{U}_n\}$ is infinite,

where $B_k = \mathbb{N}$ if $k = 0$ and $B_k = \bigcap_{m \leq k} A_m^c$ for $k > 0$.

Suppose that we have constructed $\mathcal{U}_{n_0}, \mathcal{U}_{n_1}, \dots, \mathcal{U}_{n_{k-1}}$ satisfying 1 and 2 (if $k = 0$, set $n_{-1} = -1$). Choose m and m' greater than n_{k-1} such that $B_k \in \mathcal{U}_m \cap \mathcal{U}_{m'}$ (this is possible by 2). Since \mathcal{U}_m and $\mathcal{U}_{m'}$ are distinct ultrafilters, there is an infinite $A \subseteq \mathbb{N}$ such that $A \in \mathcal{U}_m$ and $A^c \in \mathcal{U}_{m'}$. Because B_k is the disjoint union of $B_k \cap A$ and $B_k \cap A^c$, any ultrafilter \mathcal{U}_m that contains B_k as an element must contain exactly one of $B_k \cap A$ and $B_k \cap A^c$. Therefore, either $\{n \in \mathbb{N} : B_k \cap A \in \mathcal{U}_n\}$ is infinite or $\{n \in \mathbb{N} : B_k \cap A^c \in \mathcal{U}_n\}$ is infinite. If $\{n : B_k \cap A \in \mathcal{U}_n\}$ is infinite, set $n_k = m'$ and $A_k = B_k \cap A^c$. Otherwise, let $n_k = m$ and $A_k = B_k \cap A$. Clearly 1 and 2 are satisfied.

The sequence $\langle A_k \rangle$ is a pairwise disjoint family of infinite subsets of \mathbb{N} , and by construction, $A_k \in \mathcal{U}_{n_k}$. This proves (a). To prove (b), given such a sequence $\langle A_k \rangle$, we set

$$M = \bigcup_{n \in \mathbb{N}} A_{2n}.$$

Because the \mathcal{U}_n are filters, and $A_{2k} \in \mathcal{U}_{n_{2k}}$, $\{n \in \mathbb{N} : M \in \mathcal{U}_n\}$ is infinite. Similarly, for each $k \in \mathbb{N}$, $M^c \in \mathcal{U}_{n_{2k+1}}$. This means that $M \notin \mathcal{U}_{n_{2k+1}}$, so $\{n \in \mathbb{N} : M \notin \mathcal{U}_n\}$ is also infinite.

One can interpret this solution topologically. Topologize $X = \mathbb{N} \cup \{\mathcal{U}_n : n \in \mathbb{N}\}$ so that the points $n \in \mathbb{N}$ are isolated, and so that a basis for the neighborhoods of \mathcal{U}_n consists of sets of the form

$$B(A) = A \cup \{\mathcal{U}_m : A \in \mathcal{U}_m\},$$

for $A \in \mathcal{U}_n$. It is easy to see that the topology determined by this assignment of basic open sets gives a zero-dimensional Hausdorff (hence regular) topology on X . Note that the given basis consists of *clopen* (closed and open) sets. Actually, X with this topology is a subspace of the Stone-Čech compactification $\beta\mathbb{N}$ of \mathbb{N} .

Now, every infinite subset of a Hausdorff space contains an infinite relatively discrete subspace, so we can find a subsequence $\langle \mathcal{U}_{n_k} \rangle$ of \mathcal{U}_n that is relatively discrete. In a regular space, every countable relatively discrete subspace has an expansion to disjoint open sets. Thus there is a sequence of basic open sets $\langle B(A_k) \rangle$ with $A_k \in \mathcal{U}_{n_k}$ that is pairwise disjoint. In particular, these A_k are pairwise disjoint, and (a) and (b) follow.

Editorial comment. Kenneth Schilling notes that (a) cannot be improved to the existence of disjoint sets $\langle A_n \rangle$ such that each $A_n \in \mathcal{U}_n$. For a class of counterexamples, he lets $\langle B_n : n \in \mathbb{N} \setminus \{0\} \rangle$ be a sequence of disjoint nonempty subsets of \mathbb{N} , and let $\langle \mathcal{U}_n : n \in \mathbb{N} \setminus \{0\} \rangle$ be a sequence of distinct ultrafilters on \mathbb{N} such that $B_n \in \mathcal{U}_n$ for all n . Let \mathcal{V} be any nonprincipal ultrafilter on \mathbb{N} and define the ultrafilter \mathcal{U}_0 by putting $S \in \mathcal{U}_0$ if $\{n \in \mathbb{N} : S \in \mathcal{U}_n\} \in \mathcal{V}$. Then $\langle \mathcal{U}_n : n \in \mathbb{N} \rangle$ is still a sequence of distinct ultrafilters (since $B_n \notin \mathcal{U}_0$ for all n), but $S \in \mathcal{U}_0$ is also in many \mathcal{U}_n and, hence, has a nonempty intersection with B_n .

Solved also by R. J. Chapman (U. K.), J.-C. Evard, R. Holzinger, D. W. Jakel, Z. Lipecki (Poland), O. P. Lossers (The Netherlands), R. Martin (student), N. Passell, M. Scheepers, K. Schilling, L. Wertheim (student, Russia), GCHQ Problem Solving Group (U. K.), and the proposer.

10276[1993, 76]. *Proposed by Steven M. Gagola, Jr., Kent State University, Kent, OH.*

If

$$M = \begin{pmatrix} x & y \\ z & w \end{pmatrix},$$

define $\det M = xw - yz$ and $\text{dot} M = xz + yw$. Determine necessary and sufficient conditions on a field F , assumed to have characteristic different from 2, for the existence of quadratic forms $q_{ij} \in F[x, y, z, w]$ ($i, j \in (0, 1)$) such that $\det Q = (\det M)^2$ and $\text{dot} Q = (\text{dot} M)^2$, where

$$Q = \begin{pmatrix} q_{00} & q_{01} \\ q_{10} & q_{11} \end{pmatrix}.$$

In particular, do such forms exist when $F = \mathbb{Q}$?

Solution by Robin J. Chapman, University of Exeter, Exeter, U. K. Such forms exist if and only if either 2 or -2 is a square in F . Hence they do not occur if $F = \mathbb{Q}$. Let $\epsilon = 1$ or -1 . If 2ϵ is a square in F , let

$$Q = \begin{pmatrix} \sqrt{2\epsilon} \det M + \text{dot} M & -\epsilon \text{dot} M \\ \frac{1}{2} \text{dot} M & \frac{1}{2} (\epsilon \sqrt{2\epsilon} \det M - \epsilon \text{dot} M) \end{pmatrix}.$$

The verifications are immediate.

Conversely, assume that there exist quadratic forms q_{ij} with the stated properties. A straightforward computation yields, for example,

$$\begin{aligned} (q_{00}^2 + q_{01}^2)(q_{10}^2 + q_{11}^2) &= (\det Q)^2 + (\text{dot} Q)^2 \\ &= (\det M)^4 + (\text{dot} M)^4 \\ &= x^4(z^4 + w^4) + 4x^3yzw(z^2 - w^2) + 12(xyzw)^2 \\ &\quad + 4xy^3zw(w^2 - z^2) + y^4(w^4 + z^4) \\ &= f(x, y, z, w). \end{aligned}$$

Thus, f is the product of two quartic forms over F . Since the coefficients of x^4 and x^3 in f have no common factor, any nontrivial factor of f must involve x . Furthermore, $f(x, 0, z, w) = x^4(z^4 + w^4)$ must factor into two quartics each involving x , so that $z^4 + w^4$ cannot be irreducible over F . Consequently, $t^4 + 1$ is reducible over F .

If α is a root of $t^4 + 1$ in F , then $\alpha^2 = i$ (where, as usual, $i^2 = -1$); it follows that $(\alpha + \alpha^{-1})^2 = 2$, and so we are done. The only other possibility is that $t^4 + 1$ splits into two quadratic factors in F . These two factors must be among

$$\left\{ t^2 \pm i, t^2 \pm \sqrt{2}t + 1, t^2 \pm \sqrt{-2}t - 1 \right\},$$

since the roots of f in an algebraic closure of F are the primitive eighth roots of unity $(\pm 1 \pm i) / \sqrt{2}$. Thus, F contains at least one of $\{\sqrt{2}, \sqrt{-2}, i\}$. If it contains one of the first two, we are done; otherwise,

$$f(x, y, z, w) = (q_{00} + iq_{01})(q_{00} - iq_{01})(q_{10} + iq_{11})(q_{10} - iq_{11})$$

splits into four quadratic factors, and so $x^4(z^4 + w^4)$ must split into four factors each involving x . But this means that $t^4 + 1$ has a root in F . As we saw before, this implies that $\sqrt{2} \in F$, and we are done.

Solved also by the proposer.

On a Conjecture of Sophie Germain

10277[1993, 76]. Proposed by L.E. Mattics, University of South Alabama, Mobile, AL.

Let p be a prime with $p \equiv 1 \pmod{4}$. Show that there are integers x and y such that $x^p + y^p$ is of the form $u^2 + pv^2$ for integers u and v , but $x + y$ is not of that form.

Solution by A.N. 't Woord, Eindhoven University of Technology, Eindhoven, The Netherlands. We prove the claim without the condition that p be prime. Let $p > 1$ be an integer with $p \equiv 1 \pmod{4}$. Put $x = (1 + 4p)(1 + 2^p)$ and $y = 2x$. Now

$$\begin{aligned} x^p + y^p &= (1 + 2^p)x^p = (1 + 2^p)(1 + 4p)^p(1 + 2^p)^p \\ &= (1 + 4p)(1 + 4p)^{p-1}(1 + 2^p)^{p+1} \\ &= [(1 + 4p)^{(p-1)/2}(1 + 2^p)^{(p+1)/2}]^2 + p[2(1 + 4p)^{(p-1)/2}(1 + 2^p)^{(p+1)/2}]^2. \end{aligned}$$

Hence $x^p + y^p$ has the form $u^2 + pv^2$. Since squares are congruent to 0 or 1 modulo 4, and $p \equiv 1 \pmod{4}$, no number of the form $u^2 + pv^2$ can be congruent to 3 modulo 4. However,

$$x + y = 3x = 3(1 + 4p)(1 + 2^p) \equiv 3 \cdot 1 \cdot 1 = 3 \pmod{4}.$$

Hence $x + y$ does not have the form $u^2 + pv^2$.

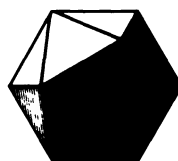
Solved also by the proposer. One incomplete solution was received.

Collaborating editors: David F. Appleyard, Paul T. Bateman, Duane M. Broline, Barry W. Brunson, Frank S. Cater, Gulbank D. Chakerian, Underwood Dudley, Gerald A. Edgar, Michael A. Filaseta, Ira M. Gessel, Richard A. Gibbs, Jerrold R. Griggs, Douglas A. Hensley, John R. Isbell, Mourad E. H. Ismail, Murray Klamkin, Daniel J. Kleitman, Frederick W. Luttman, Frank B. Miles, Richard Pfeifer, Stephen L. Portnoy, J. O. Shallit, John Henry Steelman, Kenneth B. Stolarsky, David E. Tepper, Douglas B. Tyler, Daniel Ullman, and William E. Watkins.

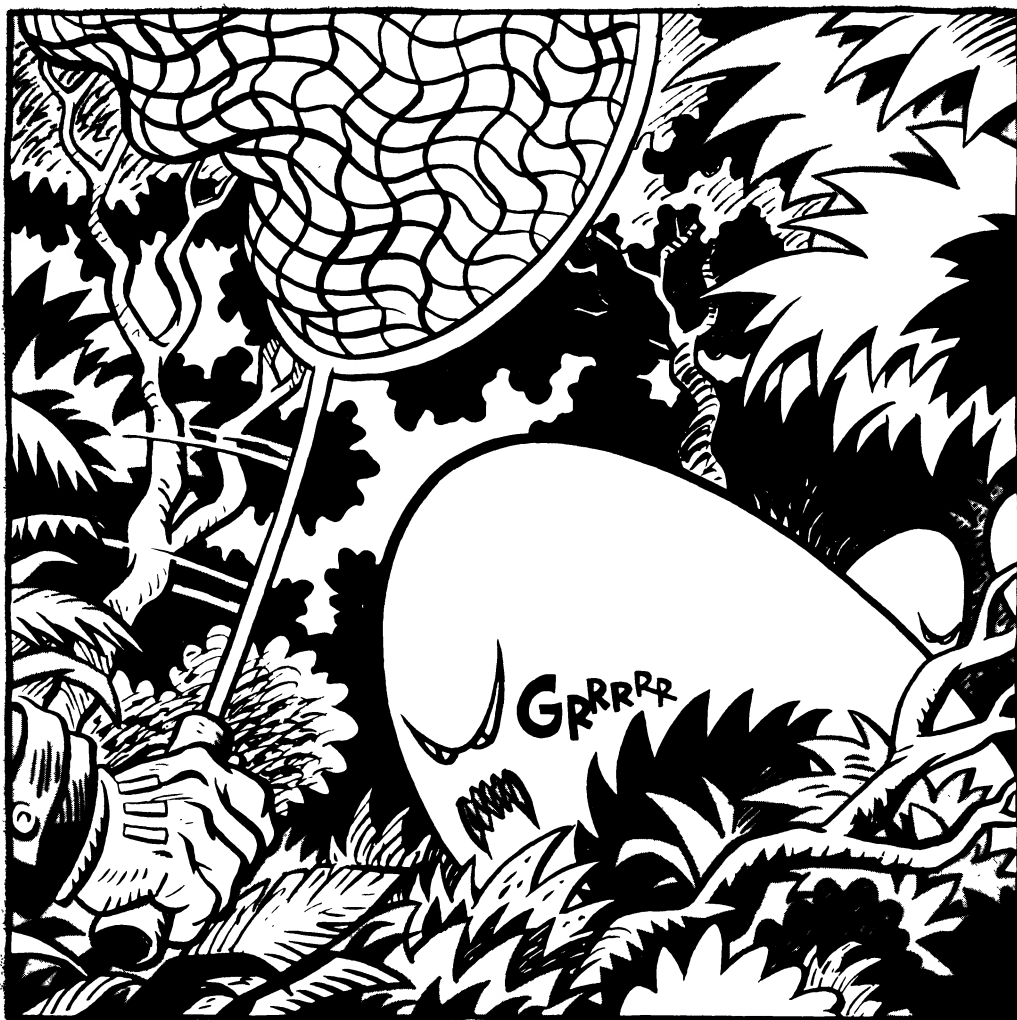
One picture is worth a thousand words, provided one uses another thousand words to justify the picture.

—H. M. Stark

The American Mathematical Monthly



Volume 102, Number 9 / NOVEMBER 1995



Stalking the Wild Ellipse
(See page 782)

NOTICE TO AUTHORS

The *Monthly* publishes articles, notes, and other features about mathematics and the profession. The readership of the *Monthly* is intended to include everybody who is mathematically inclined, including of course professional mathematicians and students of mathematics at all collegiate levels. While no single article or feature is likely to appeal to everyone, material should interest and be accessible to a large number of readers. This is the most important criterion for acceptance.

Articles may be expositions of old results or presentations of new ones. They may concern all of mathematics or one small area, a broad development or a single application, historical reminiscences or one important event. While some articles may contain the author's new research, the novelty of material and generality of the results is far less important than the clarity of exposition and general interest. Discussing one illuminating case of a well known result is far better than providing all the details of an obscure but new proposition. Articles in the *Monthly* are supposed to inform and to entertain; they are meant to be read rather than archived.

Notes are short and possibly informal articles. A note may concern a clever new proof of an old theorem, a novel way to present tired material, or a lively discussion of a philosophical (but still mathematical) issue. Also, any topic is suitable, so long as it is related to mathematics. Because a note is short, the first few sentences are the most important part: They should explain the purpose and invite the reader in. Photographs or diagrams often will attract the reader's attention.

All articles and notes should be sent to the editor:

ROGER HORN
1515 Mineral Square, Room 142
University of Utah
Salt Lake City, UT 84112

Please send 3 copies, typewritten on only one side of the paper. Illustrations should be carefully drawn on separate sheets of paper in black ink; the original should be without lettering and two copies should have appropriate captions and lettering indicated.

Proposed problems or solutions should be sent to:

RICHARD BUMBY,
P.O. Box 10971
New Brunswick, NJ 08906-0971.

Please send 2 copies of all material, typewritten if possible.

Letters to the Editor, both for publication and for private reading, should be sent to the Editor at the address given above. Comments, including criticisms, are welcome, as are all suggestions for making the *Monthly* a lively, entertaining, and informative journal.

EDITOR:

JOHN H. EWING

ASSOCIATE EDITORS:

PETER BORWEIN	FRED KOCHMAN
RICHARD BUMBY	CATHERINE MCGEOCH
DENNIS DETURCK	RICHARD NOWAKOWSKI
UNDERWOOD DUDLEY	ARNOLD OSTEBEE
JOHN DUNCAN	LEE RUBEL
JOAN FERRINI-MUNDY	ABE SHENITZER
JOSEPH GALLIAN	LYNN STEEN
STEVEN GALOVICH	STAN WAGON
RICHARD GUY	DOUGLAS WEST
DARRELL HAILE	HERBERT WILF
PAUL HALMOS	SANDY ZABELL
JOAN HUTCHINSON	PAUL ZORN

EDITORIAL ASSISTANT:

MISTY CUMMINGS

STAFF ARTIST:

MIKE CAGLE

QUOTE MASTER:

MARK WOODARD

Reprint permission:

MARCIA P. SWARD, Executive Director

Advertising Correspondence:

Ms. ELAINE PEDREIRA, Advertising Manager

Subscription correspondence, change of address, and other inquiries:

Membership / Subscriptions Department

All at the address:

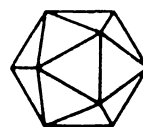
The Mathematical Association of America
1529 Eighteenth Street, N.W.
Washington, DC 20036.

Microfilm Editions: University Microfilms International, Serial Bid coordinator, 300 North Zeeb Road, Ann Arbor, MI 48106.

The AMERICAN MATHEMATICAL MONTHLY (ISSN 0002-9890) is published monthly except bimonthly June-July and August-September by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, DC 20036 and Montpelier, VT. Copyrighted by the Mathematical Association of America (Incorporated), 1995, including rights to this journal issue as a whole and, except where otherwise noted, rights to each individual contribution. General permission is granted to Institutional Members of the MAA for noncommercial reproduction in limited quantities of individual articles (in whole or in part) provided a complete reference is made to the source. Second class postage paid at Washington, DC, and additional mailing offices. **Postmaster:** Send address changes to the American Mathematical Monthly, Membership / Subscription Department, MAA, 1529 Eighteenth Street, N.W., Washington, DC, 20036-1385.

**The American
Mathematical Monthly**

Volume 102 Number 9 / NOVEMBER 1995
(ISSN 0002-9890)



Contents

ARTICLES

- My Favorite Elliptic Curve: A Tale of Two Types of Triangles /
RICHARD K. GUY 771
- Stalking the Wild Ellipse / KEITH M. KENDIG 782
- The Role of Transitivity in Devaney's Definition of Chaos /
ANNALISA CRANNELL 788
- Harvard Calculus at Oklahoma State University / KERRY JOHNSON 794
- The Stochastic Group / DAVID G. POOLE 798
- A Story of Binomial Coefficients and Primes / J. W. SANDER 802
- Turán's Graph Theorem / MARTIN AIGNER 808

FEATURES

COMMENTS 770

NOTES

- More on Kummer's Test / HANS SAMELSON 817
- The Derivative of the Exponential Map of Matrices /
G. M. TUNYMAN 818
- The Kantorovich Inequality / VLASSTIMIL PTÁK 820
- On the Generalized Inverse Form of the Equations of Constrained
Motion / ROBERT KALABA AND RONG XU 821

THE COMPUTER SCIENCE SAMPLER

- Off to the Races / JEFFREY ONDICH 826

THE EVOLUTION OF ...

- Elliptic Curves / JOHN STILLWELL 831

THE AUTHORS 838

PROBLEMS AND SOLUTIONS 840

REVIEWS

- Essays in Humanistic Mathematics*. Edited by Alvin White /
ERIC LIVINGSTON 846

TELEGRAPHIC REVIEWS 850

My Favorite Elliptic Curve: A Tale of Two Types of Triangles

Richard K. Guy

One of the many beauties of elliptic curves is their blend of arithmetic and geometry, not only intrinsically but also in their applications. If you want to learn more about them there are several good introductions available: Silverman & Tate [9], Knapp [7] and Cassels [2], who manages to write a whole book on elliptic curves without using the word 'rank.'

The curve of the title (88A in [1] or [4]) is:

$$Y^2 = X^3 - 4X + 4$$

Figure 1 shows a picture of part of its part. It's fairly uncomplicated curve: it has only one real component and doesn't break up into an 'egg' and an infinite branch as many elliptic curves do. Moreover, it doesn't have any **torsion points**, points of finite order, except for the point at infinity, which we must always remember. And I thank the referee for reminding me that when I say 'torsion points' this is an ellipsis for '**rational** torsion points.' For example, the points of inflexion are of order three, but they are not rational on this curve. One of the difficulties for the beginner is keeping track of what field he is working in: it is often convenient to vary the focus from complex to real to rational, and even to consider finite fields.

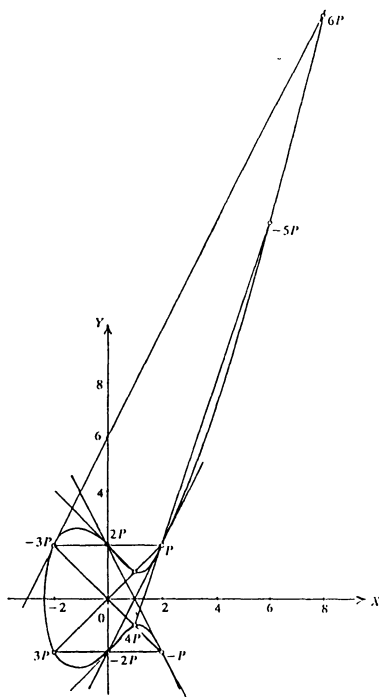


Figure 1. The elliptic curve $Y^2 = X^3 - 4X + 4$.

The curve does have several obvious rational points.

$$(0, \pm 2), \quad (1, \pm 1), \quad (2, \pm 2), \quad (-2, \pm 2).$$

The points of an elliptic curve form a group. Take the point at infinity as the (additive) identity, 0. The group law is described by noting that a straight line meets a cubic curve in three points whose sum we define to be 0. For example, the ordinate $X = 2$ meets the curve in $(2, \pm 2)$ and the point at infinity, so if $(2, \pm 2)$ are P and Q , then

$$P + Q + 0 = 0$$

and $Q = -P$. The tangent at $(2, 2)$ meets the curve again at $(0, -2) = R$, say, so that

$$P + P + R = 0,$$

$R = -2P$ and $(0, 2) = 2P$. On joining this to P we see that $(-2, 2) = -3P$, $(-2, -2) = 3P$. By joining $-P$ to $-3P$ or drawing the tangent at $-2P$ we discover that $4P = (1, -1)$ and then $5P = (6, -14)$, $6P = (8, 22)$ and so on. We soon convince ourselves that there is an infinity of rational points on the curve. In fact a theorem of Mazur (see [6], p. 223, Theorem 7.5, for example) tells us that there can't be more than 16 rational points of finite order. The **rank** of the curve is 1; all rational points can be derived from the **generator** $P = (2, 2)$.

Warning: to *prove* that a point is a generator usually requires more sophistication than we display here.

A mixture of cevians. Problem E3434 in the April 1991 MONTHLY asked, or should have asked, for integer triangle ABC in which the median from A , the bisector of angle B , and the altitude from C are concurrent. At the time of writing, no solution has been published, though I have seen an interesting one due to J. G. Mauldon, which makes no explicit use of an elliptic curve.

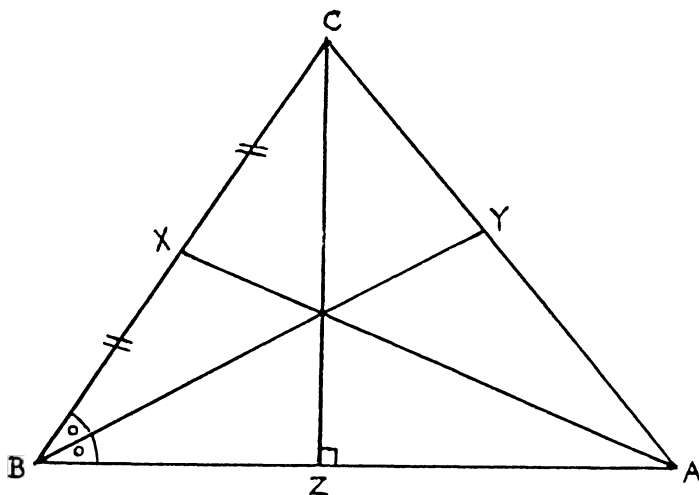


Figure 2. Triangle with concurrent median, angle-bisector and altitude.

Ceva's theorem ([3, p.4] for example) tells us that three concurrent lines drawn from the vertices of a triangle divide the sides in ratios whose product is 1:

$$\frac{BX}{XC} \cdot \frac{CY}{YA} \cdot \frac{AZ}{ZB} = 1, \quad \frac{a/2}{a/2} \cdot \frac{a}{c} \cdot \frac{b \cos A}{a \cos B} = 1$$

where the middle ratio comes from the angle-bisector theorem. Multiply $b \cos A = c \cos B$ by $2ac$ and the cosine formula gives

$$a(b^2 + c^2 - a^2) = c(c^2 + a^2 - b^2).$$

Put

$$Y = \frac{2b}{a+c}, \quad X = \frac{2c}{a+c}$$

and we get our favorite curve

$$Y^2 = X^3 - 4X + 4.$$

So we seem to have found an infinity of such triangles, but a complication is that not all rational points on the curve give real triangles. The transformation we just made inverts to

$$(a:b:c) = (2-X:Y:X).$$

We can change the signs of all three of a , b and c , so we do this if necessary to make a positive. We can change the sign of Y , since the curve is symmetrical, and so make b positive. And we can interpret either sign for c : when c is negative, Y divides CA externally in the ratio $a:c$ and BY is the *external* bisector of angle B (Figure 3).

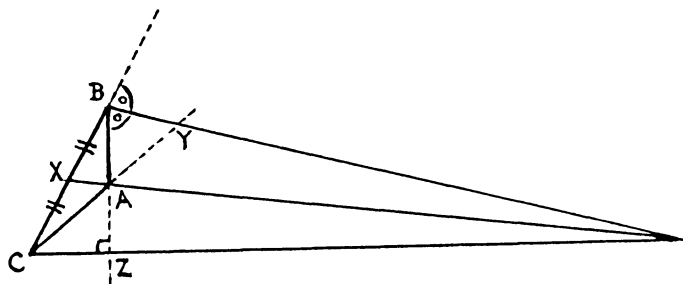


Figure 3. Triangle with external angle-bisector concurring with median and altitude.

If $X > 0$ the triangle inequality requires that $Y > 2X - 2$, $Y > 2 - 2X$ and $2 > Y$, i.e. that we are inside the region in Figure 1 bounded by the tangents at $\pm P$ and the line $Y = 2$, i.e. on the piece of curve $0 < X < 2, 0 < Y < 2$. Such points give us genuine internal bisector triangles. The point $-4P$ corresponds to the equilateral triangle.

If $X < 0$ the triangle inequality gives $Y > -2$, $Y > 2$ and $2 - 2X > Y$. We are on the piece of the curve below the tangent at $-P$ and above the line $Y = 2$: i.e. $-2 < X < 0, Y > 2$. These points give triangles whose external angle-bisector concurs with the median and altitude.

If X is outside the interval $[-2, 2]$, the triangle inequality is not satisfied. Table 1 lists a point, chosen so that $(2 - X)Y$ is positive, from each of the first twenty pairs; together with the associated triple (a, b, c) and a description of the resulting

TABLE 1. Points on curve and corresponding triangles.

point	(X, Y)	(a, b, c)	
P	$(2, 2)$	$(0, 1, 1)$	D
$2P$	$(0, 2)$	$(1, 1, 0)$	D
$-3P$	$(-2, 2)$	$(2, 1, -1)$	D
$-4P$	$(1, 1)$	$(1, 1, 1)$	$G(\Delta)$
$5P$	$(6, -14)$	$(2, 7, -3)$	N
$-6P$	$(8, -22)$	$(3, 11, -4)$	N
$7P$	$(10/9, 26/27)$	$(12, 13, 15)$	G
$8P$	$(-7/4, 19/8)$	$(30, 19, -14)$	A
$-9P$	$(-6/25, 278/125)$	$(140, 139, -15)$	A
$-10P$	$(88/49, 554/343)$	$(35, 277, 308)$	G
$11P$	$(310, -5458)$	$(308, 5458, -310)$	N
$-12P$	$(273/11^2, -3383/11^3)$	$(341, 3383, -3003)$	N
$13P$	$(206/31^2, 52894/31^3)$	$(26598, 26447, 3193)$	G
$-14P$	$(-3344/39^2, 87326/39^3)$	$(124527, 43663, -65208)$	N
$-15P$	$(9362/103^2, 1175566/103^3)$	$(610584, 587783, 4832143)$	G
$16P$	$(27105/76^2, -4131247/76^3)$	$(1182028, 4131217, -2059980)$	N
$-17P$	$(256882/151^2, -128313838/151^3)$	$(31903280, 128313838, -38789182)$	N
$18P$	$(589456/695^2, 324783646/695^3)$	$(130866415, 162391823, 204835960)$	G
$19P$	$(-2280402/1247^2, 5023772066/1247^3)$	$(3360926870, 2511886033, -1421830647)$	A
$-20P$	$(-1896655/1939^2, 17691806567/1939^3)$	$(18257812083, 17691806567, -3677614045)$	A

triangle, if any: D means degenerate, G is good, N does not yield a real triangle, while A means that the angle-bisector is external.

The point $11P$ is a pleasant surprise, though it would be natural to join $5P$ to $6P$ if one were looking for large integer points. Note that there can only be a finite number of integer points, i.e., points with integer coordinates. This is Siegel's theorem [see 8, p. 247, Theorem 3.1, for example]. Fortunately for us, any rational point will do, because the determination of all integer points requires some ingenuity, Tzanakis & Weger [13, 14] have made some progress with this problem; Zagier's paper [15] explains the connexion with the magic number g that we'll meet below. Indeed, since this paper was first drafted, a method using these **elliptic logarithms** has been developed by Stroeker & Tzanakis [11] (and independently by Gebel, Pethö & Zimmer [5]) and used by Stroeker & de Weger [12] to settle the problem of the Ochoa curve [6].

As $11P$ is quite near infinity, 11 serves as an almost period, with $12P$ near P , $13P$ near $2P$, etc. so that one can predict that (for some distance), $4P, 7P, 10P, 13P, 15P, 18P, 21P, 24P, 26P, 29P, 32P, \dots$ will give good triangles, and that $8P, 9P, 19P, 20P, 30P, 31P, \dots$ will give external bisector ones, although eventually there will be a hiccup, when a better approximation to the period takes over. About $4/11$ of the points give genuine triangles, and about $2/11$ give triangles in which it is the external bisector which concurs with the median and altitude. If you want better approximations to these fractions, or want to know just when the hiccup occurs, read on.

The 'near periods' are associated with 'large' points, such as

$$72P = (4543.72\dots, 306279.98\dots)$$

$$227P = (6619.74\dots, -538594.19\dots)$$

$$299P = (154460.66\dots, 60705331.35\dots)$$

$$1722P = (5373628.48\dots, 12456655569.68\dots)$$

These are found from the convergents to the continued fraction of the number g , defined as

$$\frac{1}{2\Omega} \int_2^\infty \frac{dX}{Y} = 0.8193959921938194669745653771\dots$$

$$= [0, 1, 4, 1, 1, 6, 3, 1, 4, 1, 4, 1, 8, 1, 4, 1, 8, 7, 5, 14, 14, 1, 1, 1, 1, 2, \dots]$$

where 2Ω is the **real period** of the curve (see later for more detail) and the lower terminal of the integral is the X -coordinate of the generator. The convergents are

$$\frac{0}{1}, \frac{1}{1}, \frac{4}{5}, \frac{5}{6}, \frac{9}{11}, \frac{59}{72}, \frac{186}{227}, \frac{245}{299}, \frac{1166}{1423}, \frac{1411}{1722}, \frac{6810}{8311}, \frac{8221}{10033}, \frac{39694}{48443}, \frac{47915}{58476}, \dots$$

whose denominators $5, 6, 11, 72, 227, 299, \dots$ are good candidates for a ‘near period.’ The lines joining $-P$ to $11P, -72P, 227P, -299P, \dots$ are closer and closer to the vertical, so that the points $-10P, 73P, -226P, 300P, \dots$ are nearer and nearer to $P = (2, 2)$; the signs have been chosen alternately so that the X -coordinates, $1.7959\dots, 1.9423\dots, 1.9520\dots, 1.9898\dots$ are less than 2: remember that the convergents are alternately less or greater than g . Figure 4 shows part of the curve magnified to illustrate the near periodicity: note that points closest together differ by $72P$.

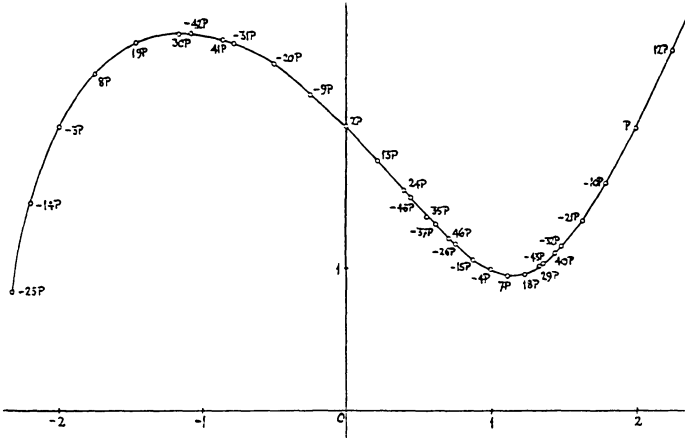


Figure 4. Curve magnified to show 11 and 72 as near periods.

An elliptic curve over the complex field should be thought of as a torus, with the real part as a circle, compactified by the point at infinity. There’s a second circle if the curve has an ‘egg.’ Figure 5 is a diagrammatic representation of the first 25 pairs of points $\pm kP$ whose labels are outside the circle and the fractional part of $kg, kg - [kg]$, is written inside the circle. The X -coordinate increases across the horizontal diameter on some curious scale, presumably related to the Weierstrass p -function. The regions of Figure 5 are labelled with the letters from the last column of Table 1. The ordinates $x = -2, 0$ and 2 give degenerate triangles, D , and the ordinate $x = 1$ corresponds to the equilateral triangle, Δ .

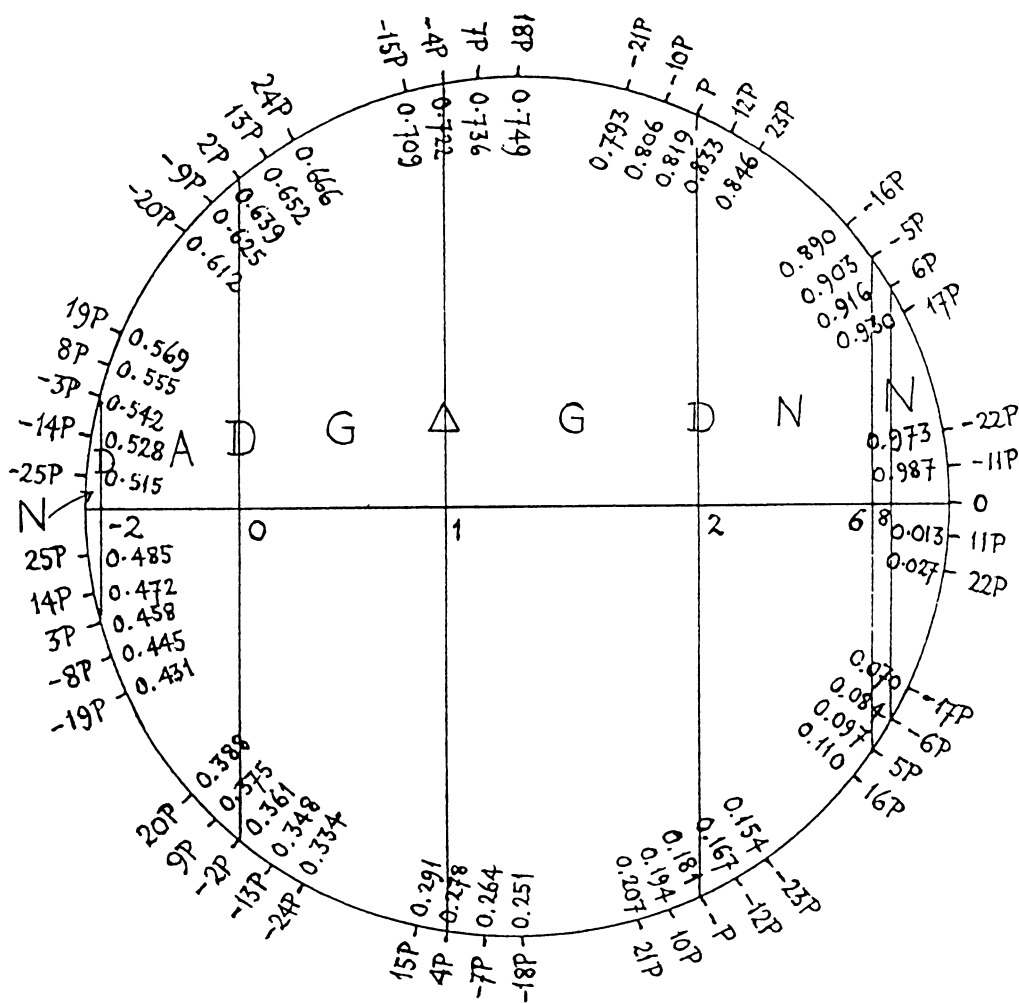


Figure 5. Diagrammatic representation showing near periodicity.

Of course, we've now given away our secret. You will have noticed that as your calculate successive points, the numbers of digits in their coordinates increase in size rather alarmingly. But the magic number g will tell us just where any point kP is: calculate the fractional part of kg and look at Figure 5. For example, $73P$ gives a good triangle (whose sides have about 140 decimal digits!) but $84P$ does not; $70P$ is good, but $81P$ requires the external bisector interpretation, as do $74P$ and $75P$.

Let \hat{E} denote the set of **real** solutions (a, b) to the equation $Y^2 = X^3 - 4X + 4$ together with the point at infinity. The real period is defined by the integral.

$$2\Omega = \int_{\alpha}^{\infty} \frac{dX}{Y} = \int_{\alpha}^{\infty} \frac{dX}{\sqrt{X^3 - 4X + 4}}$$

where α is the real root of $X^3 - 4X + 4$. Then it is true (but not so easy to prove) that there is a group isomorphism

$$\phi: E \rightarrow \frac{\mathbb{R}}{2\Omega\mathbb{Z}} \quad (a, b) \mapsto \int_a^{\infty} \frac{dX}{Y}$$

Thus the magic number g is really $g = \phi(P)$, and this explains exactly why kg being near to $2\Omega\mathbb{Z}$ is equivalent to kP being close to the point at infinity.

Here are the fractional parts of kg for the best candidates:

k	72	227	299	1423	1722	8311
	0.9965	0.0029	0.999402	0.000497	0.999899	0.000091

Problem for experts. Good approximations to a continued fraction come from truncating it just before a large partial quotient. Our continued fraction doesn't display any spectacular partial quotients, but those for several curves do. For example, curve 37A, $y(y+1) = x(x^2-1)$, has, for the magic number associated with its generator, $(0,0)$:

[0; 3, 4, 1, 1, 5, 2, **168**, **46793**, 1, 7, 1, 51, 1, 7, 1, 6, 2, 1, 1, 1, 10, 1, 2, 10, 1, 2, 11, 16, 3, 1, 1, 1, 1, 4, 1, 1, 3, 1, 1, 5, 5, 25, 1, 34, 10, 2, 18, 10, **585**, 1, 2, 3, 1, 1, **440**, 1, 1, 7, 2, 1, 4, 6, 16, 5, 2, 3, 2, 5, 1, 1, 77, 1, 2, 1, 1, 1, 13, 51, 3, 1, 2, 1, 4, 4, 3, 1, 10, 5, 1, 1, 1, 2, 1, 32, 8, 1, 2, 1, 4, 61, ...]

What is going on? Something akin to what is described by Stark in [10]?

Isosceles Heron triangles. Colleague Bill Sands is always looking for problems for *Crux Mathematicorum*; he asked if there were triangles with integer sides and area associated with rectangles having the same perimeter and area. There are indeed many such, but none of them right-angled, which is what he originally asked for. This last statement can be confirmed via curve 14A4, which has rank 0 and whose six torsion points yield only degenerate triangles. A discussion of the general problem may appear elsewhere; and see the last section for an introduction.

But here we find an infinite family of **isosceles** triangles. Let the equal legs be $m^2 + n^2$ and the base be $2(m^2 - n^2)$ so that the altitude in $2mn$:

$$\text{the semiperimeter} = p + q = 2m^2$$

$$\text{and the area} = pq = 2mn(m^2 - n^2)$$

where p and q are the sides of the associated rectangle. So we require that

$$(p - q)^2 = 4m^4 - 8mn(m^2 - n^2)$$

shall be a perfect square. If we write

$$X = \frac{2n}{m}, \quad Y = \frac{p - q}{m^2}$$

what do we get?

$$Y^2 = X^3 - 4X + 4.$$

This time all rational points give rational triangles which are realized geometrically, provided that when n is outside the interval $[0, m]$ we are willing to consider negative lengths and areas. In calculating the perimeters, sometimes the base of the triangle or one of the sides p, q of the rectangle must be taken as negative.

For comparison with the first family of triangles we use the same multiples of P as before, though now a change in sign of Y merely interchanges the roles of p and q . Write $X = x/d^2, Y = y/d^3$ where x, y, d are integers with $d > 0, x \perp d, y \perp d$ (that is, x and y are each prime to d). Note that x and y are not necessarily prime to one another: in fact x_k and y_k are both even unless k is a multiple of 4, when they are both odd, while d_k is odd unless k is a multiple of 8.

We have seen that (X_{k+1}, Y_{k+1}) may be found by joining (X_k, Y_k) to $P = (X_1, Y_1) = (2, 2)$:

$$X_{k+1} = \frac{2(X_k^2 - 2Y_k)}{(X_k - 2)^2} = \frac{2n_k}{m_k} \quad Y_{k+1} = \frac{4(X_k Y_k - 3X_k^2 + 6X_k - 4)}{(X_k - 2)^3}$$

$$\frac{x_{k+1}}{d_{k+1}^2} = \frac{(x_k^2 - 2y_k d_k)}{(x_k - 2d_k^2)^2} \quad \frac{m_{k+1}}{n_{k+1}} = \frac{(x_k - 2d_k)^2}{x_k^2 - 2y_k d_k}$$

We choose $m \perp n$ and $m > 0$; the g.c.d., (m_{k+1}, n_{k+1}) , of the numerator and denominator of the last fraction is $2d_{k-1}^2$, $16d_{k-1}^2$, $4d_{k-1}^2$ or $4d_{k-1}^2$ according as $k \equiv 0, 1, 2$, or $3 \pmod{4}$.

Table 2 lists information about the first 20 isosceles triangles and is parallel to Table 1. We do not list (m, n) since these are $(2d^2, x)$ or $(d^2, x/2)$ according as 4 divides k or not. If m and n are both odd, as they are when k is odd, we keep the triangle primitive by dividing all lengths by 2. The rectangle sides p and q are $2d(2d^3 \pm y)$ or have $\frac{1}{4}$ or $\frac{1}{8}$ of those values according as $4|k$, $2||k$ is odd. As they are each divisible by d , primitive rectangles can only be given by integer points, so that $k = 5$ and $k = 11$ are the only nontrivial examples.

The labels are the same as before, except that the interpretation of A is now: altitude and area are negative and the rectangle $p \times q$ has $q < 0 < p$, while N now means that the base of the triangle is negative, the altitude is positive or negative according as $n > m$ or $n < -m$, the area and p each have sign opposite to that of the altitude, and $q > 0$. The latter case is exemplified by $14P$ where the area is positive, but in calculating the perimeter of the triangle, its base must be taken as negative.

Shapes of triangle. In each problem, as the point moves on the curve, the shape of the triangle changes continuously. As the rational points are dense on the curve, we can approximate to any shape of triangle that is consistent with the geometrical properties that have been imposed.

The cevians triangle, for example, can be as near right-angled at C as we wish. Choose a point with X -coordinate as near to $\sqrt{5} - 1 = 1.236\dots$ as required. The point $18P$ gives a triangle with $A = 39.68^\circ$, $B = 52.40^\circ$, $C = 87.92^\circ$. The other angles approach 90° simultaneously, though not quite at the same speed, as the triangle degenerates when we approach $X = 0$; the $70P$ triangle has $A = 89.95^\circ$, $B = 88.32^\circ$, $C = 1.73^\circ$. In this problem the triangle can be equilateral, corresponding to the point $4P$, and points close by to the left or right give triangles with one or two angles less than 60° : $68P$: $(A, B, C) = (62.40^\circ, 60.98^\circ, 56.60^\circ)$ $76P$: $(A, B, C) = (57.51^\circ, 59.02^\circ, 63.47^\circ)$.

The Heron triangles are isosceles, so don't display such variety. They vary from degeneracy one way to the other: this incarnation of $70P$ gives base angles of 3.35° , while $73P$ corresponds to base angles of 88.325° . The vertical angle can also be as near to 90° as we wish: the points $15P$, $31P$, $41P$ and $57P$ give 84.7° , 93.8° , 86.3° and 87.9° .

The Heron triangles cannot be equilateral, but we can approximate by taking points near to the maxima and minima of the curve, $X = \pm 2/\sqrt{3} = \pm 1.1547\dots$. Already $4P$: $(5, 5, 6)$ and $7P$: $(53, 53, 56)$ are quite good. Next better is $30P$ with base angles 60.525° .

TABLE 2. Isosceles triangles ($m^2 + n^2, m^2 + n^2, 2(m^2 - n^2)$) and rectangles $p \times q$ with common perimeter and area.

k	x	y	d	altitude	equal legs	base
1	2	2	1	1	1	0
2	0	2	1	0	1	2
-3	-2	2	1	-1	1	0
-4	1	1	1	4	5	6
5	6	-14	1	3	5	-8
-6	8	-22	1	8	17	-30
7	10	26	3	45	53	56
8	-7	19	2	-112	113	30
-9	6	278	5	-75	317	616
-10	88	554	7	4312	4337	930
11	310	-5458	1	155	12013	-24024
-12	273	-3383	11	132132	133093	-31930
13	206	52894	31	98983	467065	912912
-14	-3344	87326	39	-5086224	5109025	-964286
-15	9362	1175566	103	49660729	67231321	90639120
16	27105	-4131247	76	626233920	868129729	-1202464642
-17	256882	-128313838	151	2928583241	8508488041	-15977204880
18	589456	324783646	695	284721984400	320177744609	292897113232
19	-2280402	5023772066	1247	-1773022816809	1859055655241	1117994669680
-20	-1896655	17691806567	1939	-28523574533020	60139308180389	105889415604678

k	area	rectangle (p, q)	
1	0	(1, 0)	D
2	0	(2, 0)	D
-3	0	(1, 0)	D
-4	12	(6, 2)	G
5	-12	(-3, 4)	N
-6	-120	(-10, 12)	N
7	1260	(60, 21)	G
8	-1680	(140, -12)	A
-9	-23100	(660, -35)	A
-10	2005080	(462, 4340)	G
11	-1861860	(-1364, 1365)	N
-12	-2109487380	(-15862, 13299)	N
13	45181384348	(871689, 51832)	G
-14	2452287298032	(4016298, 610584)	N
-15	2250602387559240	(86546265, 26004616)	G
16	-376512073310528320	(-494500840, 761398248)	N
-17	-23395287224795708040	(-4583904584, 5103790185)	N
18	41697123659341318400400	(346175467610, 120450833640)	G
19	-991115029206740553325560	(2775187436616, -357134446535)	A
-20	-1510172319128981992388733780	(125150833858190, -12066817875462)	A

A third manifestation. With help from Andrew Bremner we are investigating the general problem of finding triangle-rectangle pairs with common perimeter and common area.

Brahmagupta taught us that all Heron triangles are of shape

$$c(a^2 + b^2), \quad b(a^2 + c^2), \quad (b + c)(a^2 - bc),$$

which, if we take the third side as base, has altitude $2abc$, are $\Delta = abc(a + b)(a^2 - bc)$ and semiperimeter $s = a^2(b + c)$.

If the associated rectangle is $p \times q$, then we have $\Delta = pq$, $s = p + q$, and

$$(p - q)^2 = a^4(b + c)^2 - 4abc(b + c)(a^2 - bc)$$

must be a perfect square. Set $\mathcal{Y} = (p - q)/a^2(b + c)$, $\mathcal{X} = bc/a^2$, $\mathcal{Z} = a/(b + c)$ and the equation becomes

$$\mathcal{Y}^2 = 1 - 4\mathcal{X}\mathcal{Z} + 4\mathcal{X}^2\mathcal{Z}.$$

However, in order that this transformation be birational, we also require that

$$1 - 4\mathcal{X}\mathcal{Z}^2 = \left(\frac{b - c}{b + c}\right)^2 = \mathcal{W}^2$$

be a perfect square. On eliminating \mathcal{Z} ,

$$(\mathcal{Y}^2 - 1)^2 = 16\mathcal{X}^2\mathcal{X}^2(\mathcal{X} - 1)^2 = 4\mathcal{X}(\mathcal{X} - 1)^2(1 - \mathcal{W}^2)$$

we have a quintic surface [which deserves study in its own right]. It contains a dozen straight lines, two of which, $\mathcal{X} = 1$, $\mathcal{Y} = \pm 1$, are double, so that a plane through either of them, say

$$n(\mathcal{Y} - 1) = m(\mathcal{X} - 1)$$

cuts the surface in a cubic curve.

So we can find “all” triangle-rectangle pairs in the following sense. Such a pair corresponds to a rational point on the quintic surface. This determines (m, n) , the ‘slope’ of the plane through the point and the line $\mathcal{X} = 1$, $\mathcal{Y} = 1$. Elimination of \mathcal{Y} between the surface and the plane, yields, on writing $x = -m^4\mathcal{X}$, $y = 2m^4n^2\mathcal{W}$:

$$y^2 = x[x^2 + 2(m^4 - 2m^3n + 2n^4)x + m^6(m - 2n)^2],$$

an elliptic curve whose rational points give all triangle-rectangle pairs of ‘slope’ (m, n) . We are studying the range $0 < |m| \leq n \leq 50$.

The discriminant of the curve is $4m^{12}n^4(m - 2n)^4(m^4 - 2m^3n + n^4)$ and the curve is singular just if $m = 0$, $n = 0$, $m = n$ or $m = 2n$. The torsion group is $\mathbb{Z}/4\mathbb{Z}$, the points $(-m^3(m - 2n), \pm 2m^3n^2(m - 2n))$ being of order 4. However, if $m^4 - 2m^3n + n^4 = r^2$ is a perfect square, then the torsion group is $\mathbb{Z}/4\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$, there being additional points $(m^3(m - 2n), \pm 2m^4(m - n)(m - 2n))$ of order 4 and $((r - n^2)^2, 0)$ and $((r + n^2)^2, 0)$ of order 2.

When is $m^4 - 2m^3n + n^4$ a perfect square? Put

$$\frac{r}{m^2} = \frac{X}{2} - \frac{n^2}{m^2} \quad \text{and} \quad \frac{n}{m} = \frac{Y + 2}{2X}$$

and what do you get?

$$Y^2 = X^3 - 4X + 4$$

REFERENCES

1. B. J. Birch & W. Kuyk (editors), *Modular Functions of One Variable IV* (Proc. Internat. Summer Sch., Univ. Antwerp, 1973), Springer Lecture Notes in Math., **476**(1975), Table 1.
2. J. W. S. Cassels, *Lectures on Elliptic Curves*, London Math. Soc. Student Texts **24**, Cambridge Univ. Press, 1991.
3. H. S. M. Coxeter & S. L. Greitzer, *Geometry Revisited*, New Math. Library **19**, Math. Assoc. of America, 1967.
4. John E. Cremona, *Algorithms for Modular Elliptic Curves*, Cambridge Univ. Press, 1992, Table 1.

5. Josef Gebel, Attila Pethö & Horst G. Zimmer, Computing integral points on elliptic curves, *Acta Arith.*, (to appear).
6. Richard K. Guy, The Ochoa curve, *Crx Math.*, **16**(1990) 65–69.
7. Anthony W. Knapp, *Elliptic Curves*, Math. Notes **40**, Princeton Univ. Press, 1992.
8. Joseph H. Silverman, *The Arithmetic of Elliptic Curves*, Springer-Verlag New York, 1986.
9. Joseph H. Silverman & John Tate, *Rational Points on Elliptic Curves*, Springer-Verlag New York, 1992.
10. Harold M. Stark, An explanation of some exotic continued fractions found by Brillhart, in Atkin & Birch (editors), *Computers in Number Theory*, (*Proc. 2nd Atlas Sympos., Oxford* (1969)), Academic Press, London, 1971, pp. 21–35.
11. Roel J. Stroeker & Nikos Tzanakis, Solving elliptic diophantine equations by estimating linear forms in elliptic logarithms, *Acta Arith.*, **67**(1994), 177–196.
12. Roel J. Stroeker & Benne M. M. de Weger, On elliptic diophantine equations that defy Thue—the case of the Ochoa curve, Report 9437/B, Econ. Inst., Erasmus Univ. Rotterdam, 1994 *Experimental Math.* (submitted).
13. Nikos Tzanakis & Benne M. M. de Weger, On the practical solution of the Thue equation, *J. Number Theory*, **31**(1989) 99–132; *MR* **90c**:11018.
14. Nikos Tzanakis & Benne M. M. de Weger, How to explicitly solve a Thue-Mahler equation, *Composition Math.*, **84** (1992) 223–288; *MR* **93k**:11025; *corrections*, **89**(1993) 241–242.
15. Don Zagier, Large integral points on elliptic curves, *Math. Comput.*, **48**(1987) 425–436; *MR* **87k**:11062; Addendum, **51**(1988) 375; *MR* **89c**:11092.

Department of Mathematics & Statistics
The University of Calgary
Calgary, Alberta, CANADA T2N 1N4
rkg@cpsc.ucalgary.ca

I have never done anything “useful”. No discovery of mine has made, or is likely to make, directly or indirectly, for good or ill, the least difference to the amenity of the world ... Judged by all practical standards, the value of my mathematical life is nil; and outside mathematics it is trivial anyhow. I have just one chance of escaping a verdict of complete triviality, that I may be judged to have created something worth creating. And that I have created something is undeniable; the question is about its value.

—Godfrey H. Hardy (1877–1947)

A Mathematician's Apology, p. 150. Cambridge: Cambridge University Press, 1941.

Stalking the Wild Ellipse

Keith M. Kendig

It's not very well known, but the area of any ellipse $Ax^2 + Bxy + Cy^2 = 1$ is

$$\frac{2\pi i}{\sqrt{B^2 - 4AC}}.$$

The following examples reveal two different sides to this formula:

1. $x^2 + xy + y^2 = 1$ is an ellipse tilted from the horizontal; from $A = B = C = 1$ we easily find its area, $2\pi/\sqrt{3}$.
2. $x^2 + 4xy + y^2 = 1$ is a hyperbola. Its area is unbounded, yet $A = C = 1, B = 4$ produces a definite result, $\pi i/\sqrt{3}$.

What's going on? Is this last answer nonsense, or is the formula sending us signals, perhaps trying to tell us something informative? It turns out that if $B^2 - 4AC \neq 0$, then $Ax^2 + Bxy + Cy^2 = 1$ *always* defines an ellipse—somewhere—and the formula tells us its area. The key here is “somewhere”: far from being a simple beast, the ellipse has a decided preference for privacy. In fact if, in a specific sense, we “choose A, B, C at random”, the probability is over 81% that the ellipse is hidden from normal view. When he's away, then with probability 77% we'll get a hyperbola to look at. But our formula follows the ellipse wherever he goes. In this article, we hitch a ride, penetrating this private side of the ellipse, and report to you some unexpected findings.

Building a Cage. Most calculus students know the ellipse in “standard form”, gentle and well-behaved:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$$

While generations of students easily measure and photograph it, using semi-axes a and b . Its area is πab ($A = 1/a^2, B = 0, C = 1/b^2$).

Somewhat later, perhaps in a linear algebra course, the ellipse is fed a little mixed term Bxy , and at once it begins to stir, becoming more camera-shy. Students are supplied with correspondingly higher-tech gear, like eigenvectors. These can be used to build a cage around him; thus cornered, the ellipse relents and poses. But feed him more than a certain critical level of Bxy (in fact, do *anything* to the coefficients to make $B^2 - 4AC$ positive) and the ellipse goes wild, disappearing into parts unknown. We can't just take off in hot pursuit, for we need some idea of where to look. Recalling some particulars about making that cage will help us, and will also show how we came up with the area formula.

The Basic Eigenrecipe for Ellipses.

1. Find the roots λ_1, λ_2 of the quadratic equation

$$\det \begin{pmatrix} A - \lambda & \frac{B}{2} \\ \frac{B}{2} & C - \lambda \end{pmatrix} = 0. \quad (1)$$

2. For each λ_i , find a real unit vector v_i satisfying

$$\begin{pmatrix} A - \lambda_i & \frac{B}{2} \\ \frac{B}{2} & C - \lambda_i \end{pmatrix} v_i^t = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

If $\lambda_1 = \lambda_2$, choose $v_1 \perp v_2$. (If $\lambda_1 \neq \lambda_2$, orthogonality is automatic.)

3. Draw the rectangle having vertices

$$\pm \frac{1}{\sqrt{\lambda_1}} v_1 \pm \frac{1}{\sqrt{\lambda_2}} v_2;$$

it surrounds the ellipse, which has parametric equations

$$v = \frac{\cos(t)}{\sqrt{\lambda_1}} v_1 + \frac{\sin(t)}{\sqrt{\lambda_2}} v_2.$$

The ellipse's semi-axes are $1/\sqrt{\lambda_1}$, and $1/\sqrt{\lambda_2}$; its area is therefore $\pi/\sqrt{\lambda_1 \lambda_2}$. the little-known area formula follows from this, because $\lambda_1 \lambda_2$, being the product of the two roots, is just the constant term of the quadratic—that is, it's what we get by putting $\lambda = 0$ in (1). So

$$\lambda_1 \lambda_2 = \det \begin{pmatrix} A & \frac{B}{2} \\ \frac{B}{2} & C \end{pmatrix},$$

which is $AC - B^2/4$; this is almost the discriminant $B^2 - 4AC$. Rewriting slightly then gives us our formula.

Stalking the Ellipse. How does the eigenrecipe help to locate a hidden ellipse? Though we stated it for visible ellipses, let's try applying the recipe directly to our example $x^2 + 4xy + y^2 = 1$. The results are:

$$\begin{aligned} \lambda_1 &= 3; & \lambda_2 &= -1 \\ v_1 &= \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right); & v_2 &= \left(\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}} \right) \end{aligned}$$

$$\text{Rectangle vertices: } \pm(1/\sqrt{6}, 1/\sqrt{6}) \pm (-i/\sqrt{2}, i/\sqrt{2}).$$

Parametric eq'ns: $x = \cos(t)/\sqrt{6} - i \sin(t)/\sqrt{2}$; $y = \cos(t)/\sqrt{6} + i \sin(t)/\sqrt{2}$.

If following the recipe is actually valid, then these results suggest using space based on \mathbb{C} , not \mathbb{R} . In fact, our parametric equations do in fact define an ellipse in

\mathbb{C}^2 , and one can check that the equations indeed satisfy $x^2 + 4xy + y^2 = 1$. As it turns out, the eigenrecipe has pointed to the truth: no ellipse really takes \mathbb{R}^2 seriously as its world, but rather considers \mathbb{C}^2 as its native, rightful living space.

If this is so, what does everything look like in \mathbb{C}^2 ? Unfortunately, most of us are adept at seeing in at most three dimensions. Those rare individuals who can visualize well in four dimensions could look at $x^2 + 4xy + y^2 = 1$ in \mathbb{C}^2 , taking x and y to be complex rather than only real. Such a person would see a very rich world; for instance $x^2 + 4xy + y^2 = 1$ defines a (real, $2 - d$) surface there, and the ellipse would be seen, comfortably sitting in that surface. What's the surface look like? How is the ellipse contained in it? Where does the original hyperbola fit in?

Though we may not have $4 - d$ eyes, we do have brains capable of concocting, on occasion, admirably clever schemes. We are going to try using some strategy, together with our $3 - d$ skills, to help answer those questions and to see the good sense to the imaginary answer. Our chances of succeeding are improved if we use something in standard form, so let's try this:

$$\text{Understand } x^2 - y^2 = 1.$$

Here $A = 1, B = 0, C = -1$, so apparently the area of some ellipse is πi .

Let us write our complex variables x and y as $x = x_1 + ix_2, y = y_1 + iy_2$. Then $x^2 - y^2 = 1$ becomes

$$(x_1 + ix_2)^2 - (y_1 + iy_2)^2 = 1.$$

Note that our original picture corresponds to $x_2 = 0$ and $y_2 = 0$. Now expanding and equating real and imaginary parts gives

$$\left. \begin{aligned} x_1^2 - x_2^2 - y_1^2 + y_2^2 &= 1, \\ x_1x_2 - y_1y_2 &= 0. \end{aligned} \right\}. \quad (2)$$

Our surface in \mathbb{R}^4 is the common solution set of these two equations. If we take $3 - d$ slices of this, we'll usually see a space curve. For us, the slice $x_2 = 0$ will be a fortunate choice.

To see the part of the locus within this $3 - d$ slice, put $x_2 = 0$ in (2). We get

$$x_1^2 - y_1^2 + y_2^2 = 1, \quad y_1y_2 = 0.$$

The second equation simplifies things: $y_1y_2 = 0$ implies that either $y_1 = 0$ or $y_2 = 0$, so every point in the locus is either in the (x_1, iy_2) -plane (corresponding to $y_1 = 0$) or in the (x_1, y_1) -plane (when $y_2 = 0$). (For brevity, we'll call these planes $\mathbb{R}_{x_1y_2}$ and $\mathbb{R}_{x_1y_1}$, respectively.) If $y_2 = 0$, we get $x_1^2 - y_1^2 = 1$ which of course is what we originally had. If $y_1 = 0$, then we have $x_1^2 + y_2^2 = 1$. Here's a sketch of the part of the locus within this $3 - d$ slice:

The part in $\mathbb{R}_{x_1y_1}$ is our original hyperbola. The part in $\mathbb{R}_{x_1y_2}$ is our coveted glimpse, and it's a circle there. Now the unit of measure in the x_1 -axis is 1, and in the iy_2 -axis it is i ; thus the unit of area in $\mathbb{R}_{x_1y_2}$, the (x_1, iy_2) -plane, is $1 \cdot i = i$. In this plane, our circle is an ellipse having semi-axes 1 and i , so it has area $\pi ab = \pi i$. This is just what our area formula gives. In \mathbb{C}^2 , the eigenspaces corresponding to $\lambda_1 = +1$ and $\lambda_2 = -1$ are \mathbb{C}_x and \mathbb{C}_y , respectively. In Figure 1 we can view all of \mathbb{C}_y , so when the eigenrecipe says to draw the rectangle with vertices $(\pm 1, \pm i)$, we can really do it: the square lies in the (x_1, iy_2) -plane, neatly containing the circle there.

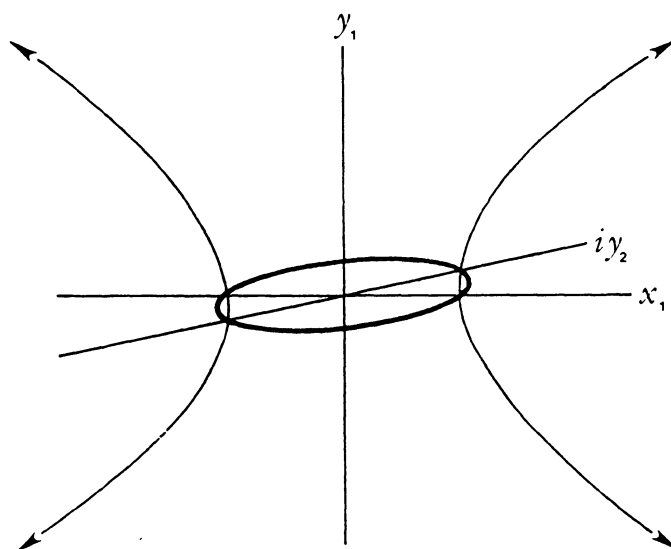


Figure 1.

A picture, basically the same as Figure 1, is obtained for $x^2 + 4xy + y^2 = 1$ in much the same way: a 45° rotation of coordinates in $\mathbb{R}_{x_1y_1}$ puts the equation into standard form; the slice $x_2 + y_2 = 0$ then does for this example what $x_2 = 0$ just did for $x^2 - y^2 = 1$.

What's the Surface? Topologically, it's trying to be a sphere. For any $Ax^2 + Bxy + Cy^2 = 1$ ($B^2 - 4AC \neq 0$), there is some linear combination f of x_1, x_2, y_1, y_2 so that the slice $f = 0$ is the union of an ellipse and a hyperbola, looking essentially like Figure 1. Topologically, the ellipse can be thought of as the "equator", and the two branches of the hyperbola as two lines of longitude crossing diametrically opposite points on that equator. The two branches approach the north and south poles. We may add those poles as two "points at infinity", completing the surface to a sphere. The two circles (equatorial and longitudinal) divide the sphere into four quarters. Within each $3 - d$ slice given by $f = a$ positive constant, the curves all have two separate branches. In the sphere, the two "ends" of any such branch meet at either the north or the south pole. Figure 2 shows how these branches fill in two diametrically opposite quarters; curves corresponding to $f = a$ negative constant similarly fill in the other two quarters. Only at $f = 0$ do we get a curve in a $3 - d$ slice with a finite loop.

Great Escapes. Our first two examples $x^2 + xy + y^2 = 1$ and $x^2 + 4xy + y^2 = 1$ show different sides of the area formula, but increasing xy to $4xy$ leads from one to the other, the first ellipse disappearing from normal view in the process. What's the journey like? The kind of transition $x^2 + Bxy + y^2 = 1$ makes as B goes from 1 to 4 is also shown in this especially simple example:

Track $x^2 + Cy^2 = 1$ as C varies from $+1$ to -1 .

The process takes place in the slice $x_2 = 0$ —in particular, in $\mathbb{R}_{x_1y_1}$ and $\mathbb{R}_{x_1y_2}$. As C decreases to 0, the circle in $\mathbb{R}_{x_1y_1}$ stretches vertically becoming, by the time C reaches 0, the two vertical lines $x_1^2 = 1$. But a little miracle always occurs precisely

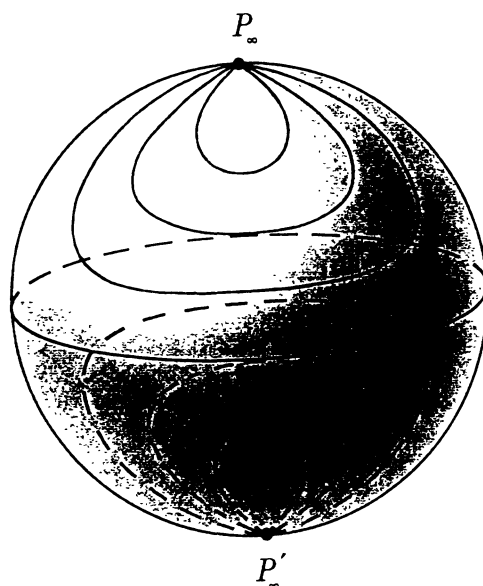


Figure 2.

at the stage when $B^2 - 4AC = 0$: the pictures in the first and second planes are congruent (and, in addition, they're both ϕ or degenerate). This acts as a kind of bridge connecting the pictures as $B^2 - 4AC$ changes sign. In our case, within the slice $x_2 = 0$ the locus of $x^2 = 1$ consists of two parallel lines $x_1^2 = 1$ in $\mathbb{R}_{x_1y_1}$, together with two parallel lines $x_1^2 = 1$ in $\mathbb{R}_{x_1y_2}$. As C decreases from 0, these last two lines close in to make an ellipse, playing in reverse what just happened in $\mathbb{R}_{x_1y_1}$.

So far we've trained our eye solely on the ellipse—that's what the formula does—but one can profitably take broader perspectives. For instance, the whole surface varies as C varies, and one can follow this along, at least topologically. There's also a middle-ground perspective that can be enlightening: If $x = x_1 + ix_2$ and $y = y_1 + iy_2$ are coordinates in eigenspaces \mathbb{C}_x and \mathbb{C}_y , the four canonical slices $\mathbb{R}_{x_iy_j}$ have a story to tell of their own. For one thing, whenever $B^2 - 4AC \neq 0$, looking at what's in each of the four slices $\mathbb{R}_{x_iy_j}$ always reveals one ellipse, two hyperbolas, and ϕ . Another fact: as $B^2 - 4AC$ changes sign, the $\mathbb{R}_{x_iy_j}$ always pair up and exchange views. In our last example for instance, ellipses and hyperbolas exchange in $\mathbb{R}_{x_1y_1}$ and $\mathbb{R}_{x_2y_2}$; in the pair $\mathbb{R}_{x_2y_1}$ and $\mathbb{R}_{x_1y_2}$, the other hyperbolas exchange with ϕ .

These facts deserve another example; let's consider $x^2 + y^2 = k$ as k goes from $+1$ to -1 (so in this case A and C grow without bound). When $k > 0$, we get circles in $\mathbb{R}_{x_1y_1}$ and ϕ in $\mathbb{R}_{x_2y_2}$. When $k = 0$, the circle has become a point, which we also see in $\mathbb{R}_{x_2y_2}$. (This of course dovetails with "area = 0" in our formula.) For $k < 0$, views exchange: the circles are in $\mathbb{R}_{x_2y_2}$ and we see nothing in $\mathbb{R}_{x_1y_1}$. In $\mathbb{R}_{x_1y_2}$ and $\mathbb{R}_{x_2y_1}$, hyperbolas exchange. When $k = 0$, one sees two crossing lines in each of $\mathbb{R}_{x_1y_2}$ and $\mathbb{R}_{x_2y_1}$.

Finally, one can follow the topological changes in the surface. In the last example, for instance, the surface for $x^2 + y^2 = k$ is a sphere if $k \neq 0$, and the equator squeezes to a point as k becomes 0. A sphere with equator pinched to a

point is topologically two spheres touching at one point. Algebraically, $x^2 + y^2 - k$ breaks up into linear factors precisely when $k = 0$; the algebraic and topological pictures thus reflect each other. Also, the longitudes (hyperbola), which form a topological loop when $k = 0$, pinch to a figure “8” when $k = 0$ —those are the two lines $x_1 + iy_2 = 0$ and $x_1 - iy_2 = 0$; each line closes to a loop at infinity, and the two loops meet at the origin.

The interested reader may wish to contemplate a few other questions:

1. In general, when should we take the negative radical in our area formula?
2. Generalize the eigenrecipe so it works for all four canonical views. (One can use $1/\sqrt{\pm \lambda_i}$ in place of $1/\sqrt{\lambda_i}$, getting four congruent rectangles.)
3. As an application of the above question, find the parametric equations (involving hyperbolic functions) for “the other hyperbola” defined by $x^2 + 4xy + y^2 = 1$. What simple change of variable transforms these to the parametric equations we met earlier for the ellipse? (Each set of equations describes the entire surface as t varies over \mathbb{C} .)
4. Trace the topological history of the surface defined by $x^2 + Cy^2 = 1$ as C varies from $+1$ to -1 .
5. For fixed A_0 and C_0 , $A_0x^2 + Bxy + C_0y^2 = 1$ describes a family of loci. Our formula gives the same result for B as for $-B$, raising the possibility that the corresponding loci might themselves be the “same” in some sense. Is there a simple linear isometry, depending only on A_0 and C_0 which for each B maps the loci $A_0x^2 \pm Bxy + C_0y^2 = 1$ into each other?
6. Let’s define “choosing A, B, C at random” to mean picking an arbitrary point in a coordinate box in (A, B, C) -space, centered at the origin. A point corresponds to an ellipse exactly when $A > 0$, $C > 0$, and $B^2 - 4AC < 0$. Show that the proportion of this box corresponding to an ellipse in the (x, y) -plane is

$$\frac{31 - 3 \ln 4}{144} \cong .1864.$$

Notice that since $B^2 - 4AC$ is homogeneous, $B^2 - 4AC = 0$ is a cone, and therefore this number is independent of the size of the box.

For further reading on viewing conics and other elementary curves in their natural habitat of \mathbb{C}^2 , and the surprises one is likely to meet, see [1] (Chapters I and II) or [2].

REFERENCES

1. K. Kendig, *Elementary Algebraic Geometry* (GTM#44), Springer-Verlag, New York, 1977.
2. K. Kendig, *Algebra, Geometry, and Algebraic Geometry: Some Interconnections*, *American Mathematical Monthly* 90 (1983) 161–173.

*Department of Mathematics,
Cleveland State University,
Euclid Avenue at East 24th street,
Cleveland, OH 44115
kendig@math.csuohio.edu*

The Role of Transitivity in Devaney's Definition of Chaos

Annalisa Crannell

1. INTRODUCTION. Devaney's definition of chaos for discrete dynamical systems is one of the most popular and most widely known. It says a function $f: M \rightarrow M$ is *chaotic* if

- (1) f is *transitive*—that is, for any pair of non-empty open sets U and V in M , there is some $k > 0$ with $f^k(U) \cap V \neq \emptyset$;
- (2) the periodic points of f are dense in M ; and
- (3) f displays the famous condition, *sensitive dependence on initial conditions*: there is a number $\delta > 0$ depending only on M and f , so that in every non-empty open subset of M one can find a pair of points whose eventual iterates under f are separated by a distance of at least δ .

Here M is generally a subset of \mathbf{R}^n , and f^n means f composed with itself n times—so that, for example, $f^3(x) = f(f(f(x)))$.

One of the ironies of this definition is that, the more popularly understood each hypothesis is, the more redundant it is in relationship to the other two.

For example, sensitive dependence is a condition which is easily understood by mathematicians and non-mathematicians alike. It has been even dubbed “the butterfly effect” in examples of popular literature such as *Jurassic Park* [3], and *The Mathematical Tourist* [7]; the phrase probably dates back to the Ray Bradbury story “A Sound of Thunder”, in which a time-traveller changes the course of history by stepping on a prehistoric butterfly [2]. This condition embodies the essence of chaos—the utter unpredictability of what ought to be simple systems—and so there is something popularly pleasing about requiring sensitive dependence on initial conditions.

However, an elegant paper by Banks, Brooks, Cairns, Davis, and Stacey [1] demonstrated that sensitive dependence is assured whenever the function displays transitivity and dense periodic points. That is, despite its popular appeal, sensitive dependence is mathematically redundant—so that in fact, chaos is a property relying only on the topological, and not on the metric, properties of a space.

The requirement that periodic points be dense is slightly less intuitive than requiring sensitive dependence, but it appeals to those who look for patterns within a seemingly random system. Mathematicians in particular instinctively seek symmetry, and the wealth of periodicities within a chaotic system is a wonderful

I am grateful to Bob Gethner, who got me interested in the questions that are asked in this paper, and to Sasha Blokh, who gave me useful and timely advice on how to pursue the answers. In addition I would like to thank Michal Misiurewicz and the reviewer for their helpful suggestions.

mathematical phenomenon. It even allows us to explain, somewhat mystically, that “there is order within chaos”. Accordingly, the search for periodic points in an understandable one.

On the other hand, Vellekoop and Berglund [8] recently gave a simple proof of an already-known theorem which says that, on any finite or infinite interval in the Real line, dense periodic points (and hence chaos) follows directly from the condition of transitivity. Moreover, they gave examples which demonstrated that neither dense periodic points nor sensitive dependence is enough to ensure any of the other conditions leading to chaos. Therefore, in one dimension both sensitive dependence and dense periodic points are redundant hypotheses in the definition of chaos.

This leaves us only the study of the transitivity hypothesis, which is required both for historical reasons and for the strength of the condition. Still, it has less intuitive justification—it is harder to explain in nonmathematical terms, and even once it is explained, it seems to follow (morally, although not mathematically) from the sensitivity hypothesis, as both of these hypotheses say that, starting with just about any data, one could eventually get just about any answer. The purpose of this paper is to ask, “why transitivity?—why not something else?” and to provide some conditions which might play the same role as transitivity, but which are slightly more intuitive.

2. A POSSIBLE ALTERNATIVE TO TRANSITIVITY. Perhaps, instead of transitivity, a more philosophically satisfying hypothesis might be one of the following:

Definition. A function $f: M \rightarrow M$ is *weakly blending* if, for any pair of non-empty open sets U and V in M , there is some $k > 0$ so that $f^k(U) \cap f^k(V) \neq \emptyset$. We say f is *strongly blending* if, for any pair of non-empty open sets U and V in M , there is some $k > 0$ so that $f^k(U) \cap f^k(V)$ contains a non-empty, open subset.

These conditions initially struck the author as an intuitive counterpart to sensitive dependence: *sensitive dependence on initial conditions* thrusts nearby points apart (for the same iterate of f), and *blending* pulls far away points together (again, for the same iterate of f)!

Blending has certain obvious disadvantages when compared with transitivity. First and foremost, any function which is blending can not be a homeomorphism, which automatically excludes the study of many interesting multi-dimensional chaotic systems—such as the horseshoe map [4, pp. 180–189]. Moreover, even in low dimensions, functions which are blending are not necessarily transitive, and transitive functions are not necessarily blending. Consider the following two examples:

Example 1. $f: S^1 \rightarrow S^1$, given by $f(\theta) = \theta + k$, where k/π is irrational. This function is rigid, irrational rotation; it is transitive but not strongly or weakly blending.

Example 2. Any continuous piecewise linear function $f: [-1, 1] \rightarrow [-1, 1]$ satisfying:

- $|f'(x)| > 2$ on except at the vertices of f ; and
- each vertex of the graph of the function lies alternately on the line $y = \pi/2$ and $y = -\pi/2$ (see the figure below).

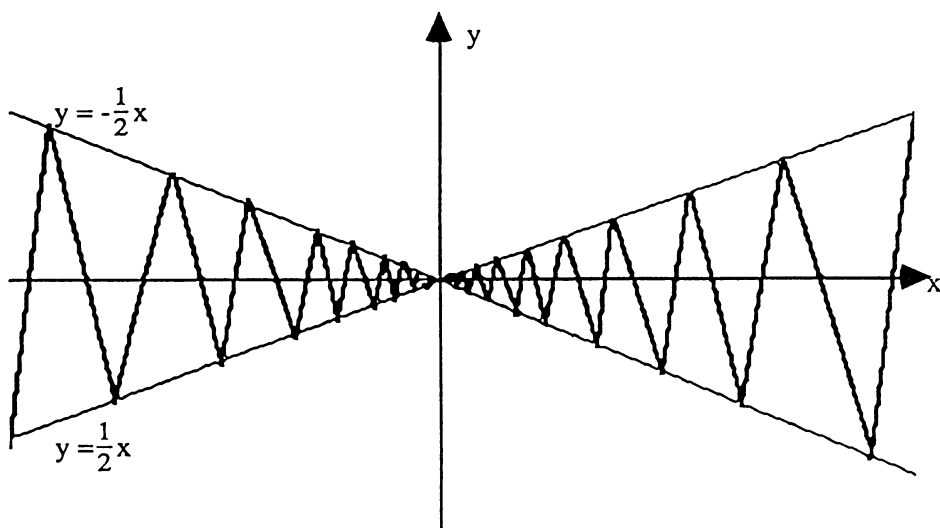


Figure 1. Graph of Example. 2

This function is clearly not transitive; in fact each set is mapped closer to the origin than it had been. At the same time, the large slope of f ensures that if neither I nor $f(I)$ contains $\{0\}$, then $f(I)$ is longer than I . This ensures that every interval is mapped, in a finite number of iterations, to an interval which contains a neighborhood of the origin—the only fixed point. Therefore, the function is strongly blending.

However, a common characteristic of these two examples is that neither has dense periodic points—in fact, the first example has no periodic points at all, and the second example has a lone fixed point. If we include dense periodic points, then the ideas of transitivity and blending in our everyday one-dimensional experience have quite a strong overlap, especially when one is considering chaos. This can be seen in theorems 1 and 2, which are the main theorems of the paper. They show that if periodic points are dense and there's a strongly repelling fixed point, then strong blending \Rightarrow transitivity \Rightarrow weak blending.

3 THE MAIN THEOREMS OF THIS PAPER. One-dimensional dynamical systems are well-understood nowadays, and so there is a wealth of theory on the subject. However, the following theorems will be proved with more simple tools: The link between open sets and continuous functions; the incredible strength of the compactness condition, and induction arguments. These simple proofs are possible because the conditions of transitivity and blending are both topological; the proofs in this section contain many of the ideas that one finds in a Point-Set Topology or an introductory Real Analysis course.

The easier of the two theorems to prove is:

Theorem 1. *Let M be a subset of \mathbf{R}^n , and $f: M \rightarrow M$ a continuous function with dense periodic points. Then if f is strongly blending, f is also transitive.*

Proof of Theorem 1. We assume that f is blending and that periodic points of f are dense. Pick two non-empty open sets, U and V . Because of the blending

property we know that there is some number $k > 0$ and some non-empty open set $N \subset M$ so that $N \subset f^k(U) \cap f^k(V)$.

For the sake of convenience of notation, we'll let $\tilde{V} = f^{-k}(N) \cap V$; \tilde{V} is the set of points in V which "blend" with those in U .

By the continuity of f , \tilde{V} is open, so our hypotheses allow us to pick a periodic point $x \in \tilde{V}$; let us say that x is of period $p > k$ (it may be that p is a multiple of the prime period of x).

Because of the way we chose $x \in \tilde{V}$, we know that $f^k(x) \in N$, and so there is some $y \in U$ with $f^k(y) = f^k(x)$. From this, the simple computation

$$f^p(y) = f^{(p-k)}(f^k(y)) = f^{(p-k)}(f^k(x)) = f^p(x) = x$$

ensures that $x \in f^p(U) \cap V \neq \emptyset$. ■

Remark. The assumption we make that N be open is a crucial one, and the theorem does not hold without it. For a counter example, consider the function

$$T(x) = \begin{cases} -(2x - 2) & \text{for } -1 \leq x \leq -\frac{1}{2} \\ 2x & \text{for } |x| < \frac{1}{2} \\ 2 - 2x & \text{for } \frac{1}{2} \leq x \leq 1 \end{cases}$$

defined on the interval $[-1, 1]$. This function is an odd extension of the tent map—its restriction to the interval $[0, 1]$ is well known to be transitive (see for example [5]). Accordingly, T has dense periodic points, and in fact every open interval in the domain eventually maps onto an interval which contains the fixed point at the origin, so that it is weakly blending. However, this function over the entire interval $[-1, 1]$ is not transitive: the interval $(0, 1)$ will never map onto any subinterval of $(-1, 0)$.

Can one hope that the converse is also true: that chaos inevitably blends all sets together (strongly)? The answer is no, unfortunately, as one can see from the following.

Example 3. We can flip the above function and get $F(x) = -T(x)$ on the interval $[-1, 1]$. This is a lovely example of a chaotic function with periodic orbits of all even periods, but no odd periods. (In fact, if x_0 is a periodic point of T with period n , then x_0 is a periodic point of F with period $2n$ —so periodic points are dense.) Examining a few iterates of this function will convince the reader that F is, moreover, transitive. On the other hand, if U is an interval to the left of the origin, and V is an interval to the right of the origin, no matter which iterate we examine we will have $F^k(U) \cap F^k(V) = \emptyset$ or $\{0\}$. Therefore, F is only weakly blending.

However, a weaker converse is true:

Theorem 2. *Let I be a compact subset of \mathbf{R} , and $f: I \rightarrow I$ a continuous, transitive function with a repelling fixed point x_0 . Then f is weakly blending.*

To prove this theorem, we will use two lemmas:

Lemma 1. *If f and x_0 are as given above, then x_0 has infinitely many eventual pre-images in I .*

Lemma 2. *If f and x_0 are as given above, then the eventual pre-images of x_0 are dense in I .*

In fact, a much stronger version of these lemmas was proved three decades ago in [6]: if f is a piecewise-monotone function, then the set

$$\{y \in I \mid f^k(y) = x \text{ for some } k\}$$

is dense in I $\forall x \in I$. As this paper needs only the weaker lemmas (with the weaker hypotheses), we will restrict our proofs accordingly.

Proof of Theorem 2. We wish to show that for any two open sets $U, V \subset I$, there is some $n > 0$ with $f^n(U) \cap f^n(V) \neq \emptyset$. Lemma 2 tells us that the eventual preimages of x_0 are dense, and so there exist $u \in U$, $v \in V$, and $j, k > 0$ so that $f^j(u) = x_0 = f^k(v)$. Assume without loss of generality that $k > j$; then because x_0 is fixed, we have $f^k(u) = x_0 = f^k(v)$. Thus, $x_0 \in f^k(U) \cap f^k(V) \neq \emptyset$, and our theorem is proved. ■

Proof of Lemma 1. We will prove this lemma by induction.

Suppose x_0 is our given repelling fixed point, and we are given a finite set $X_n = \{x_{-n}, \dots, x_{-1}, x_0\}$ with $f(x_k) = x_{k+1}$, $k = -n, \dots, -1$. If $n = 0$, then we have $X_0 = \{x_0\}$.

Choose an open set $U \subset I$ with $X_n \subset U$ satisfying.

- (1) if $y \in U$ then $f(y) \neq x_{-n}$ (unless $n = 0$ and $y = x_0$); and
- (2) $f(U \setminus B_\varepsilon) \cap B_\varepsilon = \emptyset$.

(Here B_ε is assumed to be the ball of radius ε centered at x_{-n} .) In the case $n = 0$, we use the fact that x_0 is repelling to satisfy the second of these two assumptions.

From here, we will use transitivity to show that f must send the exterior of the set U arbitrarily close to x_{-n} : that is, for every $\varepsilon > 0$, $f(U^c) \cap B_\varepsilon \neq \emptyset$.

We can choose U sufficiently small that U^c contains an open set. By the transitivity of f , we know that $f^{k+1}(U^c) \cap B_\varepsilon \neq \emptyset$ for some $k \geq 0$; we're trying to show that $k = 0$.

Let Y be the set of points which start in the complement of U and which are first mapped into B_ε on the $k + 1^{\text{st}}$ iteration. That is,

$$Y = \{y \in U^c \mid f^{k+1}(y) \in B_\varepsilon: \quad f^j(y) \notin B_\varepsilon \text{ if } 1 \leq j \leq k\}.$$

Then clearly $f^k(Y) \cap B_\varepsilon = \emptyset$. Moreover, we must have $f^k(y) \in U^c$ if $y \in Y$, for if it were otherwise, assumption (2) would give us

$$f^{k+1}(y) = f(f^k(y)) \in f(U \setminus B_\varepsilon) \subset (B_\varepsilon)^c$$

for some $y \in T$. This contradicts the definition of Y . Therefore, we see that $f^k(Y) \subset U^c$ and that $f(f^k(Y)) \cap B_\varepsilon \neq \emptyset$ —so $f^k(Y)$ is the subset of U^c which proves our claim.

The rest of the proof of Lemma 1 follows easily, for the claim holds regardless of the size of ε , and therefore the compactness of U^c tells us that there is some point y in U^c with $f(y) = x_{-n}$.

This argument gives us an infinite sequence $\{x_{-k}\}_{k=0}^\infty$ with $f^k(x_{-k}) = x_0$, and so completes the proof of Lemma 1. ■

Proof of Lemma 2. Let $X = \{y \in I \mid f^k(y) = x_0 \text{ for some } k\}$. We want to show that X is dense in I . Because f is transitive, it follows that if X is anywhere dense, then X must be everywhere dense. Let's assume that opposite: that X is totally disconnected.

If such is the case, the X^c must be open, so we can write $X^c = \bigcup_{k=1}^{\infty} I_k$, where the I_k 's are distinct, open intervals in I . Lemma 1 notes that X is infinite, so, in fact, there must be an infinite number of such intervals.

Note, moreover, that at least one interval has x_0 as an endpoint; call this interval I_1 . Because x_0 is fixed, we have $f^2(I_1) \cap I_1 \neq \emptyset$ —in fact, because of the construction of the I_k 's, we have $f^2(I_1) \subseteq I_1$.

On the other hand, transitivity prohibits exactly such a cycle, for I_1 must visit each of the infinite number of intervals—a contradiction. This contradiction arose from assuming that X is not dense in I , and so our final lemma is proved. ■

REFERENCES

1. J. Banks, J. Brooks, G. Gairns, G. David, and R. Stacey, *On Devaney's definition of chaos*, American Mathematical Monthly **99** (1992), 332–334.
2. Ray Bradbury, *A Sound of Thunder*, Golden Apples of the Sun, Doubleday & Company, Inc., 1953.
3. Michael Crichton, *Jurassic Park: a novel*, Knopf, New York, 1990.
4. Bob Devaney, *An Introduction to Chaotic Dynamical Systems*, second edition, Addison-Wesley, 1989.
5. Steven N. MacEachern and L. Mark Berliner, *Aperiodic chaotic orbits*, American Mathematical Monthly **100** (1993), 237–241.
6. William Parry, *Symbolic dynamics and transformations of the unit interval*, Transactions of the American Mathematical Society **122** (1966), 368–378.
7. Ivars Peterson, *The mathematical tourist: snapshots of modern mathematics*, Freeman, New York, 1988.
8. M. Vellekoop and R. Berglund, *On Intervals, Transitivity = Chaos*, American Mathematical Monthly **101** (1994), 353–355.

Department of Mathematics
Franklin & Marshall College
Lancaster, PA 17604
A_Crannell@ACAD.FandM.edu

This paper gives wrong solutions to trivial problems. The basic error, however, is not new.

—Clifford Truesdell
Mathematical Reviews **12**, p. 561.

Harvard Calculus at Oklahoma State University

Kerry Johnson

Calculus reform is being discussed throughout the mathematics community. Part of the reason that the reform movement has so much momentum comes from the fact that annually 300,000 students enroll in “engineering” calculus and only 140,000 finish the year with a grade of D or higher [4]. Part of the reason involves money: The National Science Foundation has awarded nearly \$11 million dollars to the reform movement [3].

The calculus reform movement began in the late-1980’s following a wave of general-education reforms [7]. Oklahoma State University jumped on the reform bandwagon in the Fall of 1992. The OSU Mathematics Department was given an NSF grant to try the calculus materials written by the Consortium based at Harvard and to disseminate the results of its efforts through conferences. Since this was experimental, the department offered two types of courses: one using the Harvard materials [1, 2] and one using a standard text [6]. This provided an opportunity to compare the two groups of calculus students.

Calculus at OSU is taught in two 15 week semesters. Calculus 1 covers one variable differentiation and integration topics and Calculus 2 deals with series and multivariate calculus.

Here are some natural questions to ask:

1. *Do Harvard students make better grades in calculus than traditional calculus students?*
2. *Are Harvard students more likely to enroll in subsequent mathematics courses?*
3. *Do Harvard students perform better in subsequent mathematics courses than other students?*
4. *How do students that go from Harvard Calculus 1 into Standard Calculus 2 perform?*

The data used in answering these questions comes from the Fall 1992 to the Spring 1994 semester. The data was taken from all students who enrolled in calculus during these semesters.

Question 1: *Do Harvard students make better grades in calculus than traditional calculus students?*

Yes. A higher percentage of the Harvard students pass the course and make a C or better in the course than traditional students. For example, 67 percent of the Harvard Calculus 1 students made a C or better in Calculus 1, while only 62 percent of the traditional students made a C or better in Calculus 1. In Calculus 2,

80 percent of the Harvard students made a C or better, where as only 71 percent of the traditional students made a C or better. In addition, these results are fairly stable because of the number of students involved (more than 110 students in each of the fur cases).

Question 2: *Are Harvard students more likely to enroll in subsequent mathematics courses?*

That depends on the students' major. For example, engineering students at OSU are required to take mathematics courses through Differential Equations. The breakdown of the students' home colleges (e.g. Engineering, Agriculture, Business, etc.) for Harvard students and traditional students turn out to be very similar however.

The largest difference in the enrollments was in Calculus 2. For students who made a D or better in Calculus 1, 63 percent of the Harvard students took Calculus 2 while only 56 percent of the traditional students took the course. However, more of the Harvard students switched to traditional Calculus 2 (44 percent) than traditional students switched to Harvard Calculus (18 percent).

Enrollments in Different Equations were similar (Harvard-36 percent, traditional-33 percent), while enrollments in Linear Algebra differed by about 7 percent (Harvard-20 percent, traditional-27 percent). These percents may change when enough time has passed for both groups to take these courses.

Question 3: *Do Harvard students perform better in subsequent mathematics courses than other students?*

The simple answer is no. Harvard students do not seem to do as well in subsequent mathematics courses as their counterparts from the traditional course. Only 45 percent of the 120 students that had a D or better in Harvard Calculus 1 were able to make the same grade or better in Calculus 2, where as 53 percent of the 227 traditional students that made a D or better in Calculus 1, maintained their grade in Calculus 2. Of these students, 70 percent of the Harvard students made a C or better in Calculus 2 while 82 percent of the traditional students made a C or better in Calculus 2. One reason for this is that several of the Harvard students took traditional Calculus 2 rather than Harvard Calculus 2. As the table under Question 4 indicates, this is not the best possible combination of courses.

The numbers are too small for higher level courses to reach any solid conclusions, but the trend seems to be the same as for Calculus 2. In Differential Equations, 50 percent of the D or better Harvard Calculus 2 students maintained or improved their grades, while 58 percent of the traditional students did. In Linear Algebra, 60 percent of the Harvard students made the same grade or better that they made in Calculus 2, while 69 percent of the traditional students maintained or improved their grades in Differential Equations. In addition, fewer Harvard students made a C or better in Differential Equations and Linear Algebra than did the traditional students. It is important to note that the comparisons in both Differential Equations and Linear Algebra involve less than twenty Harvard students and 81 traditional students.

Question 4: *How do students that go from Harvard Calculus 1 into Standard Calculus 2 perform?*

TABLE 1

Calculus 1	Calculus 2	Total number of students	% that made a C or better in Calculus 2	% that made the same grade or better in Calc 2 that they made in Calc 1
Traditional	Traditional	185	81.1	55.1
Traditional	Harvard	25	92	56
Harvard	Traditional	47	55.3	25.5
Harvard	Harvard	67	80.6	62.7

The table above summarizes the data. This only includes those students who made a D or better in Calculus 1.

Based on the above information, traditional Calculus 1 students should take Harvard Calculus 2, and the Harvard Calculus 1 students should take Harvard Calculus 2. The worst possible mixture of the above courses is Harvard Calculus 1 and traditional Calculus 2. This makes sense since traditional calculus focuses more on algebra and memorizing formulas while Harvard calculus focuses more on how these formulas can be applied to real life problems and why they are true.

Of the students who switched from the traditional course into the Harvard course, 40 percent made an A in Calculus 1, 36 percent a B, and 20 percent a C. Of the students who switched from the Harvard course into the traditional course, only 21 percent made an A in Calculus 1, while 38 percent made a B, and 36 percent made a C. Again this may be due to the algebraic focus of traditional calculus. Those students who switched from the Harvard course into the traditional course may have been good with algebra and were expecting to get an A in Calculus 1. When they discovered that this was not the calculus they learned in high school, they switched.

Conclusions: The overall trend seems to be that Harvard calculus grades are better than traditional students grades, yet traditional students tend to be more successful in subsequent math courses. This is most likely due to Harvard students moving from an application based course to a more algebraically based course such as traditional Calculus 2, Differential Equations, or Linear Algebra. It appears that Harvard calculus does address the problem of students passing calculus, but may not fully prepare students for algebraically based subsequent courses.

Another explanation for the higher grade levels in Harvard calculus is the *novelty effect*. Subjects show increased interest, motivation, or participation simply because they are doing something different [5]. Most likely, the increased grade levels are due to both the *novelty effect* and the Harvard materials. In any case, the grades in Harvard calculus do seem higher.

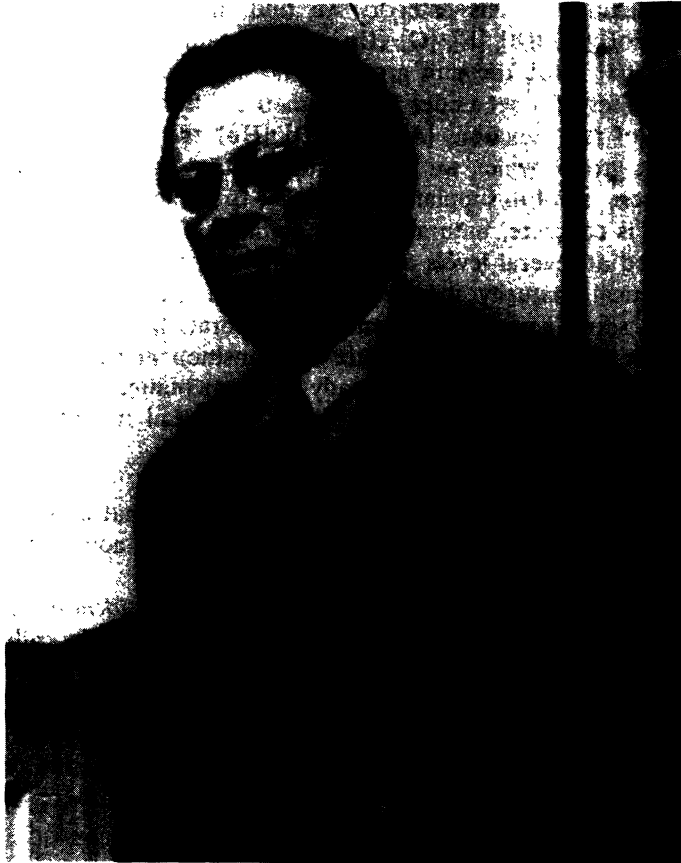
REFERENCES

1. Consortium based at Harvard, *Calculus*, John Wiley & Sons, Inc., New York, 1994.
2. Consortium based at Harvard *Multivariable Calculus* (Draft version), John Wiley & Sons, Inc., New York, 1994.
3. E. Culotta, The calculus of education reform, *Science*, 255(5048) (1992) 1060–1062.
4. J. Ferrini-Mundy and K. G. Graham, An overview of the calculus curriculum reform effort: Issues for learning, teaching, and curriculum development, *American Mathematical Monthly*, 98(7) (1991) 627–635.

5. L. R. Gay, Educational research: Competencies for analysis and application, Macmillian Publishing Company, New York, 1992.
6. R. E. Larson, R. P. Hostetler, and D. E. Heyd, Calculus with analytic geometry, D. C. Heath, Lexington, Mass., 1979.
7. C. J. Mooney, As wave of curricular reform continues, its scope and effectiveness are questioned, *Chronicle of Higher Education*, 38(18) (1992) 18, A15.

*Department of Mathematics
Oklahoma State University
401 Mathematical Sciences
Stillwater, OK 74078
kerry@math.okstate.edu*

PICTURE PUZZLE
(from the collection of Paul Halmos)



More than fifty years ago he was as young as this.
(see page 816)

The Stochastic Group

David G. Poole

Consider this to be propaganda. We want to present a “new” example of a group and lobby for its inclusion in a first course in abstract algebra. Virtually all abstract algebra textbooks include the standard examples of groups: the integers under addition, the integers modulo n under addition modulo n , $m \times n$ matrices over the reals (or the integers) under matrix addition, the group of symmetries of a regular polygon, the symmetric group, various subgroups of the multiplicative group of nonzero complex numbers, and so on. A biased sample of such books¹ reveals that, beyond these basic examples, there is not consensus on which further examples to include. In this category we find the general linear group (and subgroups thereof) [A], [BB], [E], [G], [Hu], the (one-dimensional) affine group [A], [BB], [B], [Du], [H], [Hu], the unit group of the integers modulo n [BB], [D], [G], [Hu], the power set of a set under the symmetric difference operator [BC], [Bu], various groups of real functions [A], [Bu], [D], [Hu], the Möbius group [B], and a host of abstract groups which we will attempt neither to identify nor enumerate.

Our intent here is not to supplant any of these examples but to add to the list an example which is concrete, natural (in our opinion), flexible (in the sense that it can be presented at several levels of sophistication and can grow with students as their mathematical maturity increases) and which affords an opportunity for students to do some genuine mathematical exploration. Our example is based upon the notion of the stochastic matrix (or transition matrix) associated with a Markov chain. Many students will already have encountered this example in a linear algebra course; a slight modification of the usual definition allows us to approach it from a group-theoretic point of view.

Definition. A *stochastic matrix* over a field F is a square matrix with entries from F with the property that the entries in each of its columns add up to 1.

If F is an ordered field, then we may define a *positively stochastic matrix* to be a stochastic matrix with nonnegative entries—this is what is usually called a stochastic matrix, in the case where F is the field of real numbers. We will not need this concept here.

Our interest is in the set of all nonsingular stochastic matrices over F under the operation of matrix multiplication. When F is a finite field, this produces nice examples for students to explore. For instance, over \mathbf{Z}_2 there are only two nonsingular stochastic 2×2 matrices,

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

¹The ones on my bookshelf.

and they form a cyclic group of order 2 under multiplication. Over \mathbf{Z}_3 , there are six such matrices:

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 2 & 1 \\ 2 & 0 \end{bmatrix}, \begin{bmatrix} 2 & 0 \\ 2 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 2 \\ 1 & 2 \end{bmatrix}, \begin{bmatrix} 1 & 2 \\ 0 & 2 \end{bmatrix}.$$

Again, it can easily be checked that they form a group under multiplication. After a few examples of this nature, most students will have been persuaded that there is a theorem lurking in the background and it is a nice exercise to see if they can formulate and prove it. While the 2×2 case can be done by “brute force” without too much difficulty (restrict F to be the rationals or the reals if fields are unfamiliar), most students will not want to contemplate even the 3×3 —much less the general—case. They should therefore be suitably impressed by the elegance and generality of the following proof once they have covered a few preliminary results on group actions.

Theorem A. *The set S of all nonsingular stochastic $n \times n$ matrices over a field F forms a group under matrix multiplication.*

Proof: Consider S as a subset of $GL(n, F)$, the general linear group of all nonsingular $n \times n$ matrices over F . Write the elements of F^n , where F^n denotes the vector space of all n -tuples over F , as row vectors and let $GL(n, F)$ act on F^n as matrix multiplication from the right. Let $\mathbf{j} = (1, 1, \dots, 1) \in F^n$ and recall that for any $n \times n$ matrix M over F , $\mathbf{j}M$ has the column sums of M as its coordinates. It is clear from the definition of stochastic matrix that S is the stabilizer of \mathbf{j} and hence is a subgroup of $GL(n, F)$ [G, Exercise 5.27].

Definition. The group of all nonsingular stochastic $n \times n$ matrices over a field F is called the *stochastic group* of $n \times n$ matrices over F and is denoted by $S(n, F)$. If p is prime and F is the Galois field of order $q = p^m$, we write $S(n, q)$ instead of $S(n, F)$.

The next problem which naturally arises is to determine the structure of the stochastic group—can we classify it in terms of known groups? The extent to which this can be discussed in an undergraduate course depends upon the level of the course and the background of the students. But even in an introductory abstract algebra course, much can be done. For example, it is clear that $S(2, 2) \cong \mathbf{Z}_2$ and students should have little difficulty showing that $S(2, 3) \cong S_3$, the symmetric group on three elements (take $\begin{bmatrix} 0 & 2 \\ 1 & 2 \end{bmatrix}$ and $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ as generators in the second example above). It takes a bit more effort to see that $S(2, 4) \cong A_4$ and that $S(3, 2) \cong S_4$. Advanced undergraduates might discover that $S(2, 5)$ has the presentation $\langle a, b \mid a^5 = b^4 = 1, ba = a^2b \rangle$ from which it follows that $S(2, 5) \cong \mathbf{Z}_4 \rtimes_{\theta} \mathbf{Z}_5$, the semidirect product of \mathbf{Z}_5 by \mathbf{Z}_4 with $\theta: \mathbf{Z}_4 \rightarrow \text{Aut}(\mathbf{Z}_5)$ the homomorphism determined by $\Theta_b(a) = a^2$ (take $a = \begin{bmatrix} 0 & 4 \\ 1 & 2 \end{bmatrix}$ and $b = \begin{bmatrix} 1 & 4 \\ 0 & 2 \end{bmatrix}$, for example).

But before we lose sight of the forest for the trees, let us step back from these examples and ask ourselves if there really is any pattern here. To continue with the finite stochastic groups for a moment, we might first ask for the order of $S(n, q)$. The proof of the next proposition is analogous to the corresponding derivation of the order of $GL(n, q)$ [R, Theorem 8.10]. We begin with an easy lemma.

Definition. The set of all vectors in F^n whose components add up to 1 is denoted by $U(F^n)$.

Lemma. Let F be a field, and let $\alpha_1, \alpha_2, \dots, \alpha_k \in U(F^n)$.

- (i) If $\beta = \sum_{i=1}^k c_i \alpha_i$ for some $c_i \in F$, $1 \leq i \leq k$, then $\beta \in U(F^n) \cap \text{span}\{\alpha_1, \alpha_2, \dots, \alpha_k\}$ if and only if $\sum_{i=1}^k c_i = 1$.
 (ii) If $F = GF(q)$ is the Galois field with q elements, and if $\{\alpha_1, \alpha_2, \dots, \alpha_k\}$ is linearly independent, then

$$|U(F^n) \cap \text{span}\{\alpha_1, \alpha_2, \dots, \alpha_k\}| = q^{k-1}.$$

Proof:

- (i) Let $\beta = \sum_{i=1}^k c_i \alpha_i \in \text{span}\{\alpha_1, \alpha_2, \dots, \alpha_k\}$. Let $\alpha_i = (\alpha_{i1}, \alpha_{i2}, \dots, \alpha_{in})$ for $1 \leq i \leq k$ and let $\beta = (\beta_1, \beta_2, \dots, \beta_n)$. Then

$$\sum_{i=1}^k c_i = \sum_{i=1}^k c_i \left(\sum_{j=1}^n \alpha_{ij} \right) = \sum_{j=1}^n \sum_{i=1}^k c_i \alpha_{ij} = \sum_{j=1}^n \beta_j.$$

Thus $\sum_{i=1}^k c_i = 1$ if and only if $\sum_{j=1}^n \beta_j = 1$.

- (ii) Since $\{\alpha_1, \alpha_2, \dots, \alpha_k\}$ is linearly independent, there is a one-to-one correspondence between k -tuples (c_1, c_2, \dots, c_k) and vectors $\beta = \sum_{i=1}^k c_i \alpha_i \in \text{span}\{\alpha_1, \alpha_2, \dots, \alpha_k\}$. By (i), $\beta \in U(F^n) \cap \text{span}\{\alpha_1, \alpha_2, \dots, \alpha_k\}$ if and only if $\sum_{i=1}^k c_i = 1$ so there are q^{k-1} distinct β 's (for we can choose the first $k-1$ c_j 's arbitrarily and then set $c_k = 1 - \sum_{j=1}^{k-1} c_j$).

Proposition. $|S(n, q)| = q^{n-1}(q^{n-1} - 1)(q^{n-1} - q) \cdots (q^{n-1} - q^{n-2})$.

Proof: Let $A \in S(n, q)$. Since A is nonsingular, the columns of A form an ordered basis $\{\alpha_1, \alpha_2, \dots, \alpha_n\}$ of F^n in $U(F^n)$. There are q^{n-1} choices for α_1 (the last component being determined by the criterion that the components add up to 1). There are then $q^{n-1} - 1$ choices for α_2 since we must rule out any vector in the span of α_1 (which, in this case, is just α_1 itself). Next, α_3 is any vector in $U(F^n)$ which is not the span of α_1 and α_2 . By the lemma, this means there are $q^{n-1} - q$ choices for α_3 . In general, the lemma implies that there are $q^{n-1} - q^{i-2}$ choices for α_i , $i \geq 2$. The result follows.

The problem of classifying the stochastic groups still remains. The examples given above do not suggest an obvious pattern so the following result comes as a pleasant surprise.

Theorem B. For any field F , $S(n, F) \cong \text{Aff}(n-1, F)$ for all $n \geq 2$.

Recall that the affine group $\text{Aff}(n-1, F)$ consists of all mappings $\alpha: F^{n-1} \rightarrow F^{n-1}$ of the form $\alpha(\mathbf{x}^T) = A\mathbf{x}^T + \mathbf{b}^T$, $\mathbf{x} \in F^{n-1}$, where $A \in GL(n-1, F)$ and $\mathbf{b} \in F^{n-1}$, together with the operation of composition. The monomorphism $\varphi: \text{Aff}(n-1, F) \rightarrow GL(n, F)$ defined by $\varphi(\alpha) = \begin{bmatrix} A & \mathbf{b}^T \\ \mathbf{0} & 1 \end{bmatrix}$ maps $\text{Aff}(n-1, F)$ isomorphically onto the subgroup of $GL(n, F)$ consisting of all nonsingular matrices whose last row is $[0, 0, \dots, 0, 1]$ [R]. We will identify $\text{Aff}(n-1, F)$ with this subgroup.

Proof: For $n \geq 2$, let $Q = \begin{bmatrix} I & \mathbf{0}^T \\ \mathbf{j} & 1 \end{bmatrix} \in GL(n, F)$ where I is the $(n-1) \times (n-1)$ identity matrix and $\mathbf{j} = [1, 1, \dots, 1] \in F^{n-1}$. Define $\varphi \in \text{Aut}(GL(n, F))$ to be the inner automorphism determined by Q ; that is, for $X \in GL(n, F)$, $\varphi(X) = QXQ^{-1}$. Observe that $Q^{-1} = \begin{bmatrix} I & \mathbf{0}^T \\ -\mathbf{j} & 1 \end{bmatrix}$.

We claim that $S(n, F)$ and $\text{Aff}(n-1, F)$ are in fact conjugate subgroups of $GL(n, F)$ under φ . Let $M \in S(n, F)$ and partition M as $\begin{bmatrix} N & \mathbf{u}^T \\ \mathbf{n}_0 & u_0 \end{bmatrix}$ where $N \in M_{n-1}(F)$, $\mathbf{u}, \mathbf{n}_0 \in F^{n-1}$, $u_0 \in F$ and $\mathbf{j}N + \mathbf{n}_0 = \mathbf{j}$, $\mathbf{j}\mathbf{u}^T + u_0 = 1$. Then

$$\varphi(M) = QMQ^{-1} = \begin{bmatrix} N - \mathbf{u}^T\mathbf{j} & \mathbf{u}^T \\ \mathbf{0} & 1 \end{bmatrix}$$

and, since φ is inner, $\det \varphi(M) = \det M \neq 0$. Hence $N - \mathbf{u}^T\mathbf{j} \in GL(n-1, F)$ and so $\varphi(M) \in \text{Aff}(n-1, F)$.

Conversely, if $P = \begin{bmatrix} A & \mathbf{b}^T \\ \mathbf{0} & 1 \end{bmatrix} \in \text{Aff}(n-1, F)$ then, setting $M = Q^{-1}PQ$, we find that

$$M = \begin{bmatrix} A + \mathbf{b}^T\mathbf{j} & \mathbf{b}^T \\ \mathbf{j} - \mathbf{j}A - \mathbf{j}\mathbf{b}^T\mathbf{j} & 1 - \mathbf{j}\mathbf{b}^T \end{bmatrix}.$$

Again, $\det M = \det P \neq 0$ and, checking the column sums, we find that

$$\mathbf{j}(A + \mathbf{b}^T\mathbf{j}) + (\mathbf{j} - \mathbf{j}A - \mathbf{j}\mathbf{b}^T\mathbf{j}) = \mathbf{j} \text{ and } \mathbf{j}\mathbf{b}^T + (1 - \mathbf{j}\mathbf{b}^T) = 1.$$

Thus M is stochastic and, since $\varphi(M) = P$, it follows that φ maps $S(n, F)$ isomorphically onto $\text{Aff}(n-1, F)$ as claimed.

One of the nicest advantages of this representation of the affine group is that subgroups are easy to find, or perhaps we should say easy to name: just add the adjective “stochastic” to the name of any subgroup of the general linear group. Thus, we obtain the *special stochastic group* (stochastic matrices with determinant 1), the *upper triangular stochastic group*, and so on. We might also consider the *doubly stochastic group* consisting of all nonsingular stochastic matrices whose rows also sum to 1. The possibilities are endless; hopefully, enterprising readers will decide to continue this investigation.

ACKNOWLEDGMENT. I would like to thank the referee for his helpful comments concerning this paper. In particular, I am indebted to him for suggesting an improved version of the Lemma.

REFERENCES

- [A] R. B. J. T. Allenby, *Rings, Fields and Groups: An Introduction to Abstract Algebra* (2nd edition), Edward Arnold, London, 1991.
- [BB] J. A. Beachy and W. D. Blair, *Abstract Algebra with a Concrete Introduction*, Prentice-Hall, Englewood Cliffs, NJ, 1990.
- [BC] D. C. Buchthal and D. E. Cameron, *Modern Abstract Algebra*, PWS Publishers, Boston, 1987.
- [B] R. P. Burn, *Groups: A Path to Geometry*, Cambridge University Press, Cambridge, 1985.
- [Bu] D. M. Burton, *Abstract Algebra*, Wm. C. Brown, Dubuque, 1988.
- [D] R. A. Dean, *Classical Abstract Algebra*, Harper & Row, New York, 1990.
- [Du] J. R. Durbin, *Modern Algebra: An Introduction* (3rd edition), John Wiley & Sons, New York, 1992.
- [E] G. Ehrlich, *Fundamental Concepts of Abstract Algebra*, PWS-Kent, Boston, 1991.
- [G] J. A. Gallian, *Contemporary Abstract Algebra* (2nd edition), D. C. Heath, Lexington, MA, 1990.
- [H] I. N. Herstein, *Abstract Algebra* (2nd edition), Macmillan, New York, 1990.
- [Hu] T. W. Hungerford, *Abstract Algebra: An Introduction*, Saunders, Philadelphia, 1990.
- [R] J. J. Rotman, *An Introduction to the Theory of Groups* (3rd edition), Allyn & Bacon, Boston, 1984.

Department of Mathematics
Trent University Peterborough,
Ontario Canada K9J 7B8
madgp@blaze.trentu.ca

A Story of Binomial Coefficients and Primes

J. W. Sander

1. INTRODUCTION. Primes, the atoms of the integers, naturally have a long history in mathematics. Around 1800, *Gauss and Legendre* made rather precise conjectures about the asymptotic behavior of the function $\pi(x)$ counting the primes up to x . The first mathematician who proved any worthwhile results about this behavior was *Chebyshev* in 1851/52 (see [1], paragraph 7). He showed that

$$C_1 \frac{x}{\log x} < \pi(x) < C_2 \frac{x}{\log x} \quad (1)$$

for some constants $0 < C_1 < 1 < C_2$. It is well known that in fact

$$\pi(x) = \frac{x}{\log x} + r(x),$$

where the error term $r(x)$ is of smaller order than the main term $x/\log x$. This result, the so-called prime number theorem, was proved independently by *Hadamard* and *de la Vallée-Poussin* at the end of the 19th century using complex analysis. The bounds in (1), however, can be deduced in a very elementary way. Surprisingly (at first), the middle binomial coefficients $\binom{2n}{n}$ are an appropriate tool. Why is this so?

Let us write down $\binom{2n}{n}$ in its prime factor decomposition for a few small (but not too small) values of n :

$$\begin{aligned} \binom{120}{60} &= 2^4 \cdot 3^2 \cdot 7 \cdot 13 \cdot 17 \cdot 23 \cdot 31 \cdot 37 \cdot 61 \cdot 67 \cdot 71 \cdot 73 \cdot 79 \cdot 83 \cdot 89 \cdot 97 \cdot 101 \cdot 103 \\ &\quad \cdot 107 \cdot 109 \cdot 113 \end{aligned}$$

$$\binom{122}{61} = 2^5 \cdot 3^2 \cdot 7 \cdot 11^2 \cdot 13 \cdot 17 \cdot 23 \cdot 31 \cdot 37 \cdot 61 \cdot \dots \cdot 113$$

$$\binom{124}{62} = 2^5 \cdot 3^3 \cdot 7 \cdot 11^2 \cdot 13 \cdot 17 \cdot 23 \cdot 31 \cdot 37 \cdot 67 \cdot \dots \cdot 113$$

$$\binom{126}{63} = 2^6 \cdot 3 \cdot 5^3 \cdot 11^2 \cdot 13 \cdot 17 \cdot 19 \cdot 23 \cdot 37 \cdot 41 \cdot 67 \cdot \dots \cdot 113$$

$$\binom{128}{64} = 2 \cdot 3 \cdot 5^3 \cdot 11^2 \cdot 13 \cdot 17 \cdot 23 \cdot 37 \cdot 41 \cdot 67 \cdot \dots \cdot 113 \cdot 127$$

The dots indicate the presence of all the primes in the given range.

The first observation is that these integers have many prime factors and at the same time are “almost” squarefree, where “squarefree” means that an integer has no repeated prime factor. A closer look reveals that every prime p , $n < p \leq 2n$, divides $\binom{2n}{n}$ exactly once. This is clear, since each such p divides the numerator

$(2n)!$ of $\binom{2n}{n}$ once, but cannot divide $n!$ and therefore is no factor of the denominator. One may also notice that the primes p , $2n/3 < p \leq n = 2n/2$, do not divide $\binom{2n}{n}$ (this is because p and $2p$, but not $3p$, are factors of $(2n)!$, and p divides $n!$), and this pattern continues.

The main idea for proving *Chebyshev's* result (1) now is to use $\binom{2n}{n}$ as an approximations for $\prod_{p < 2n} p$, where p runs through the sequence of primes. The size of $\binom{2n}{n}$ in turn can be estimated by Stirling's well-known formula for factorials. Taking logarithms and replacing $2n$ by x , we obtain

$$x \approx \log \prod_{p \leq x} p = \sum_{p \leq x} \log p \approx \pi(x) \log x. \quad (2)$$

The last asymptotic equality in (2) is derived by partial summation, the simple sum-analogue of partial integration.

One may ask for the advantage of $\binom{2n}{n}$ over $\prod_{p \leq 2n} p$. The answer is that binomial coefficients, in addition to their multiplicative properties (which reflect their mimicking products of primes), also have a nice additive property, namely their recursion formula. This provides the chance to use induction, which actually is done in the course of the proof.

Paul Erdős certainly has been and still is asking the most questions in number theory (and other areas of mathematics as well as life in general). He came across the beautiful idea to use binomial coefficients for a proof of *Chebyshev's* theorem in the 1930's. The importance of this result is but one reason why one likes to know as much as possible about the multiplicative structure of binomial coefficients. Therefore, it is no surprise that *Erdős* conjectured that for $n > 4$, the integers $\binom{2n}{n}$ are never squarefree, although they seem to be "almost squarefree". He even asked the more general question: Given a positive integer a , do we always find some prime p such that $p^a \mid \binom{2n}{n}$, if n is sufficiently large?

In 1985, *Sárközy* [10] proved that $\binom{2n}{n}$ is never squarefree for all sufficiently large $n \geq n_0$, thus answering *Erdős's* first question in the affirmative for large n . In 1988, *Goetgheluck* ([6], [7]) gave a numerical verification of this conjecture for all $n \leq 2^{42205184}$; we remark that if n is not a power of 2, then $4 \mid \binom{2n}{n}$ (as the reader may verify by use of Lemma 1(i) below); since 4 is a square, we only have to take care of those n type $n = 2^k$. Recently *Velammal* [11] proved *Sárközy's* theorem with an explicit constant $n_0 = 2^{8000}$. Independently, a similar result was obtained by *Granville* and *Ramaré* [8]. By checking the finitely many $\binom{2n}{n}$ for all $n = 2^k$, $2 < k < 8000$, this confirms *Erdős's* first conjecture. The general question has been answered in the affirmative by the author [9] in 1992. These results depend on estimates of so-called exponential sums, a deep and useful tool in analytic number theory.

In this note, we shall deal with these problems in a purely elementary manner, obtaining, however, weaker results. We define for a positive integer a and a prime p

$$E_{a,p}(N) = \text{card} \left\{ n : 0 \leq n < N, p^a \nmid \binom{2n}{n} \right\}$$

and

$$E_a(N) = \text{card} \left\{ n : 0 \leq n < N, p^a \nmid \binom{2n}{n} \text{ for all primes } p \right\}.$$

$E_a(N)$ is the number of exceptions n , $0 \leq n < N$, to *Erdős'* conjecture for prime powers p^a . The results of *Sárközy* [10] and the author [9] show that $E_a(N)$ is bounded for every fixed a , i.e. the exceptional set of *Erdős'* problems is always finite.

Our elementary method leads to

Theorem. (i) For integers $a \geq 1$ and $k \geq 0$,

$$E_{a,2}(2^k) = \sum_{t=0}^{a-1} \binom{k}{t},$$

i.e., in the set

$$\left\{ \binom{0}{0}, \binom{2}{1}, \binom{4}{2}, \dots, \binom{2(2^k-1)}{2^k-1} \right\}$$

of the first 2^k middle binomial coefficients, there are exactly $\sum_{t=0}^{a-1} \binom{k}{t}$ elements not divisible by 2^a .

(ii) For integers $a \geq 1$, $k \geq 0$, and a prime $p \geq 3$,

$$E_{a,p}(p^k) \geq \left(\frac{p+1}{2} \right)^k \sum_{t=0}^{a-1} \binom{k}{t}.$$

Note that if $a > k$, then $\sum_{t=0}^{a-1} \binom{k}{t} = 2^k$, which is the cardinality of the set in (i). Therefore, 2^a divides none of the coefficients in the set.

Theorem 2. (i) For any integer $a \geq 1$, we have

$$E_{a,2}(N) < C_a (\log N)^{a-1},$$

where C_a is a constant depending only on a .

(ii) Let $a \geq 1$ be an integer and p a prime. For any $\epsilon > 0$ and sufficiently large N ,

$$E_{a,p}(N) \leq N^{\gamma_p + \epsilon},$$

where

$$\gamma_p = \frac{\log \frac{p+1}{2}}{\log p}.$$

For any prime p and $a \geq 1$, we may choose ϵ small enough that $\gamma_p + \epsilon < 1$. Therefore, (ii) immediately implies the following

Corollary. Let $a \geq 1$ be an integer and p a prime. Then

$$\lim_{N \rightarrow \infty} \frac{E_{a,p}(N)}{N} = 0.$$

We say that $E_{a,p}$ has asymptotic density 0. Since obviously $E_{a,p}(N) \geq E_a(N)$, we conclude that the exceptional set in *Erdős'* problem has asymptotic density 0 as well.

For references on this subject and related questions see [2], [3], [4] and [5].

2 PROOF OF THE THEOREMS. There is a simple formula for calculating the exponent $e_p(n!)$ of the prime p in $n!$, namely

$$e_p(n!) = \frac{n - S_p(n)}{p - 1}, \quad (3)$$

where $S_p(n)$ denotes the sum of the digits of n written in base p , i.e. $S_p(n) = \sum n_i$, where

$$n = n_s p^s + n_{s-1} p^{s-1} + \dots + n_1 p + n_0 \quad (4)$$

for some integer s and $0 \leq n_i < p$, $n_s > 0$. The easy induction proof for (3) uses the following argument: Let p^j be the exact power of p dividing n . Then $e_p(n!) = e_p((n-1)!) + j$, and the last j p -ary digits of n are 0's. By subtracting 1, these last j digits turn into $(p-1)$'s, and the digit before them is decreased by 1. The corresponding formula $S_p(n) = S_p(n-1) - j(p-1) + 1$ yields the induction step.

From this one obtains the charming fact that $e_p\left(\binom{m+n}{n}\right)$ is exactly the number of "carries" occurring while adding m and n in p -ary notation (4).

For a prime p and a positive integer n , let $L_p(n)$ be the number of digits $n_i \geq p/2$ in the above p -ary expansion. The following lemma which has been well known to *Legendre* and *Kummer* shows that the order of the prime p in $\binom{2n}{n}$ is closely related to the number of "large" digits in the p -ary representation of n .

Lemma 1. (i) For a positive integer n ,

$$e_2\left(\binom{2n}{n}\right) = L_2(n) = S_2(n).$$

(ii) For a positive integer n and a prime p ,

$$e_p\left(\binom{2n}{n}\right) \geq L_p(n).$$

Proof: (i) If n is written in base 2, a multiplication of n by 2 simply means a shift of digits, thus $S_2(2n) = S_2(n)$. By (3), we get

$$e_2\left(\binom{2n}{n}\right) = e_2((2n)!) - 2e_2(n!) = S_2(n).$$

(ii) Let $2n = \sum n'_i p^i$ be the p -ary expansion of $2n$. For $n_i \geq p/2$, we clearly have $n'_i \leq 2n_i - p + 1$. Thus by (3)

$$e_p\left(\binom{2n}{n}\right) = \frac{1}{p-1}(2S_p(n) - S_p(2n)) = \frac{1}{p-1} \sum_i (2n_i - n'_i) \geq L_p(n).$$

The next lemma shows that a slowly increasing sequence (b_N) of integers with $b_N \approx N$ for infinitely many N satisfies $b_N \approx N$ for all N .

Lemma 2. Let $(b_N)_{N \geq 1}$ be a sequence of positive integers satisfying

$$b_{N+1} - b_N \in \{0, 1\} \text{ for } N \geq 1. \quad (5)$$

Furthermore, let $(m_k)_{k \geq 1}$ be a strictly increasing sequence of positive integers. Define integers r_k as the difference of b_{m_k} and m_k by

$$b_{m_k} = m_k + r_k \text{ for } k \geq 1. \quad (6)$$

Then we have for $m_k \leq N < m_{k+1}$

$$|b_N - M| \leq \max(|r_k|, |r_{k+1}|).$$

Proof: By (5) and (6),

$$b_{m_k} \leq b_N \leq b_{m_k} + (N - m_k) = N + r_k.$$

Also

$$b_{m_{k+1}} \geq b_N \geq b_{m_{k+1}} - (m_{k+1} - N) = N + r_{k+1}.$$

Together, we get

$$N + r_{k+1} \leq b_N \leq N + r_k,$$

which proves the lemma.

Proof of Theorem 1. We want to count the number of n , $0 \leq n < N$, such that $2^a \nmid \binom{2n}{n}$ for some fixed a . By Lemma 1 we may count as well the number of n which have at most $a - 1$ digits 1 in their binary representation. If N is a power of 2, this is a simple combinatorial exercise. For general prime powers, $N = p^k$ is the easiest case, but we only obtain upper bounds according to Lemma 1(ii).

(i) By Lemma 1(i)

$$\begin{aligned} E_{a,2}(2^k) &= \text{card}\{n: 0 \leq n < 2^k, L_2(n) < a\} \\ &= \text{card}\{(n_0, \dots, n_{k-1}) \in \{0, 1\}^k: \sum n_i < a\} \\ &= \sum_{t=0}^{a-1} \text{card}\{(n_0, \dots, n_{k-1}) \in \{0, 1\}^k: \sum n_i = t\} \\ &= \sum_{t=0}^{a-1} \binom{k}{t}. \end{aligned}$$

(ii) Since there are $(p - 1)/2$ p -ary digits $\geq p/2$, and these are counted in $L_p(n)$, we get by Lemma 1(ii)

$$\begin{aligned} E_{a,p}(p^k) &= \text{card}\left\{n: 0 \leq n < p^k, e_p\left(\binom{2n}{n}\right) < a\right\} \\ &\leq \text{card}\{n: 0 \leq n < p^k, L_p(n) < a\} \\ &= \sum_{t=0}^{a-1} \text{card}\{n: 0 \leq n < p^k, L_p(n) = t\} \\ &= \sum_{t=0}^{a-1} \binom{k}{t} \left(\frac{p-1}{2}\right)^t \left(\frac{p-1}{2}\right)^{k-t} \\ &\leq \left(\frac{p+1}{2}\right)^k \sum_{t=0}^{a-1} \binom{k}{t}. \end{aligned}$$

Proof of Theorem 2. It was convenient to calculate $E_{a,p}(N)$ for $N = p^k$. For general N , it is almost impossible to determine the exact value of $E_{a,p}(N)$. For this reason, we have to be satisfied with an approximation, obtainable by Lemma 2, since $E_{a,p}$ is indeed growing slowly and is well-estimated for all $N = p^k$.

Let

$$b_N = b_N(a, p) = \text{card}\left\{n: 0 \leq n < N, p^a \mid \binom{2n}{n}\right\} = N - E_{a,p}(N).$$

(i) By Theorem 1(i) for large k , and using the crude upper bound $\binom{k}{t} \leq k^t$,

$$|b_{2^k} - 2^k| = E_{a,2}(2^k) = \sum_{t=0}^{a-1} \binom{k}{t} \leq ak^{a-1}.$$

Setting $m_k = 2^k$, we obtain by Lemma 2 for $2^k \leq N < 2^{k+1}$ and a suitable constant C_a

$$E_{a,2}(N) = |b_N - N| \leq a(k+1)^{a-1} \leq a(\log_2 N + 1)^{a-1} < C_a(\log N)^{a-1}.$$

(ii) For general p , Theorem 1(ii) yields in a similar fashion for $p^k \leq N < p^{k+1}$

$$E_{a,p}(N) \leq \left(\frac{p+1}{2}\right)^{k+1} a(k+1)^{a-1}. \quad (7)$$

Now we choose N resp. k large enough that

$$\log \frac{p+1}{2} + \log a + (a-1)\log(k+1) < \epsilon k \log p.$$

Adding $k \log(p+1)/2$ on both sides, we obtain

$$\begin{aligned} (k+1)\log \frac{p+1}{2} + \log a + (a-1)\log(k+1) \\ < \frac{\log \frac{p+1}{2}}{\log p} k \log p + \epsilon k \log p \\ = (\gamma_p + \epsilon)k \log p \leq (\gamma_p + \epsilon)\log N. \end{aligned}$$

Taking exponentials, we have exactly what is needed in (7) to prove the theorem.

REFERENCES

1. H. Davenport, *Multiplicative Number Theory*, 2nd edition (revised by Hugh L. Montgomery), Springer-Verlag, New York-Heidelberg-Berlin, 1980.
2. P. Erdős, *Problems and Results on Number Theoretic Properties of Consecutive Integers and Related Questions*, Proc. Fifth Manitoba Conf. on Numerical Mathematics (1975), 25–44.
3. P. Erdős, R. L. Graham, *On the prime factors of $\binom{n}{k}$* , Fibonacci Quarterly **14** (1976), 348–352.
4. P. Erdős, R. L. Graham, *Old and New Problems and Results in Combinatorial Number Theory*, Monographie No. 28 de L'Enseign. Math., Geneva, 1980.
5. P. Erdős, R. L. Graham, I. Ruzsa, E. G. Straus, *On the prime factors of $\binom{2n}{n}$* , Math. Comp. **29** (1975), 83–92.
6. P. Goetgheluck, *Computing binomial coefficients*, Amer. Math. Monthly **94** (1987), 360–365.
7. P. Goetgheluck, *On prime divisors of binomial coefficients*, Math. Comp. vol. **51**, no. 183 (1988), 325–329.
8. A. Granville, O. Ramaré, *Explicit bounds on exponential sums and the scarcity of squarefree binomial coefficients*, (to appear).
9. J. W. Sander, *Prime power divisors of binomial coefficients*, J. reine, angew. Math. **430** (1992), 1–20.
10. A. Sándor, *On Divisors of Binomial Coefficients, I*, J. Number Theory **20** (1985), 70–80.
11. A. Velammal, *Is the binomial coefficient $\binom{2n}{n}$ squarefree?*, Hardy-Ramanujan J. **18**(1995), 23–45.

Inst. f. Math. Univ. Hannover
Welfengarten 1, 30167
Hannover GERMANY
sander@math.uni-hannover.de

Turán's Graph Theorem

Martin Aigner

One of the fundamental results in graph theory is the Theorem of Turán, proved in 1941, which initiated extremal graph theory. (See the book [2] by Bollobás as a standard reference.) Turán's theorem was rediscovered many times, and it is the purpose of this article to discuss some of the most beautiful older and more recent proofs.

Let us fix some notation. We consider graphs G on the vertex-set $V = \{1, 2, \dots, n\}$ and edge-set $E \subseteq \binom{V}{2}$. If i and j are neighbors, then we write $ij \in E$. A k -clique in G is a complete subgraph of G with k vertices, denoted by K_k . Turán posed the following question: Suppose G does not contain a k -clique, how many edges can G maximally have? Let us denote this number by $t(n, k)$. We have $t(n, 2) = 0$, and $t(n, k)$ is clearly an increasing function in k .

We readily obtain examples of such graphs by dividing V into $k - 1$ pairwise disjoint subsets, $V = V_1 \cup \dots \cup V_{k-1}$, $|V_i| = n_i$, $n = n_1 + \dots + n_{k-1}$, joining two vertices if and only if they lie in distinct V_i, V_j . Let us denote the resulting graph by $K_{n_1, \dots, n_{k-1}}$. Figure 1 shows the graph $K_{2, 2, 3}$.

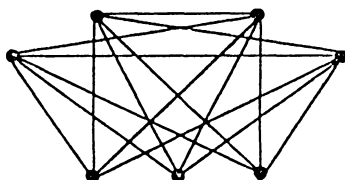


Figure 1.

The graph $K_{n_1, \dots, n_{k-1}}$ contains $\sum_{i \neq j} n_i n_j$ edges, and it is clear that we obtain a maximal number of edges among these graphs if we divide the numbers n_i as evenly as possible, i.e. $|n_i - n_j| \leq 1$ for all i, j . If, in particular, $k - 1$ divides n , then we may choose $n_i = n/k$ for all i , obtaining

$$\binom{k-1}{2} \frac{n^2}{(k-1)^2} = \frac{k-2}{k-1} \cdot \frac{n^2}{2}$$

edges. Turán's theorem now states that this number is an upper bound for the edge-number of any graph G on n vertices without k -cliques.

Theorem of Turán. *Let $G(V, E)$ be a graph on n vertices without a k -clique, then*

$$|E| \leq \frac{(k-2)n^2}{2(k-1)}. \quad (1)$$

More precisely, the theorem states that the graph $K_{n_1, \dots, n_{k-1}}$ with $|n_i - n_j| \leq 1$ for $i \neq j$ is the *unique* graph without a k -clique with the maximal number $t(n, k)$ of edges. These graphs are therefore called *Turán graphs* $T(n, k)$. In the following, we will restrict ourselves to showing (1), but in some of the proofs we will demonstrate that the graphs $T(n, k)$ attain the maximum for arbitrary k . The uniqueness is then supplied by an easy argument.

As a warm-up let us look at the first interesting case $k = 3$: A triangle-free graph contains at most $n^2/4$ edges, and the unique extremal graph is $K_{n/2, n/2}$ if n is even, respectively $K_{(n-1)/2, (n+1)/2}$ if n is odd. For this special case, proofs were known before Turán's work. Before we look at two of them we need some more notation.

The *degree* d_i of vertex i is the number of edges incident with i . By counting in two ways we obtain

$$\sum_{i=1}^n d_i = 2|E|. \quad (2)$$

A set $A \subseteq V$ is called *independent*, if A contains no edges. As an example, all the defining vertex-sets V_i in the graph $K_{n_1, \dots, n_{k-1}}$ are independent. The number $\alpha(G) = \max(|U|: U \subseteq V \text{ independent})$ is called the *independence number* of G .

$k = 3$: **First Proof** (Mantel 1906). Let $ij \in E$. Since G contains no triangles we have $(d_i - 1) + (d_j - 1) \leq n - 2$ (see Figure 2), hence $d_i + d_j \leq n$. Summing over the edges we obtain

$$\sum_{ij \in E} (d_i + d_j) \leq n|E| \quad (3)$$

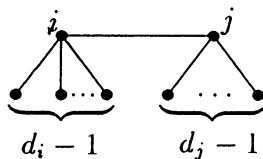


Figure 2.

The number d_i clearly appears d_i times in the sum of (3), and we conclude

$$\sum_{ij \in E} (d_i + d_j) = \sum_{i=1}^n d_i^2 \leq n|E|. \quad (4)$$

By the Cauchy-Schwarz inequality $(\sum x_i y_i)^2 \leq \sum x_i^2 \cdot \sum y_i^2$ applied to $x_i = d_i$, $y_i = 1$ we obtain by (2) and (4)

$$n^2|E| \geq \sum_{i=1}^n d_i^2 \cdot n \geq \left(\sum_{i=1}^n d_i \right)^2 = 4|E|^2, \quad (5)$$

and thus $|E| \leq n^2/4$. \square

Let us demonstrate how the uniqueness of the extremal graph $K_{n/2, n/2}$ is established for n even. (The case n odd is analogous.) If $|E| = n^2/4$, then we must have equality in (5). Now, we have equality in the Cauchy-Schwarz inequality iff the vectors are multiples of each other. For the vector (d_i) this means $d_i = d$ for all i , and we conclude $n^2|E| = n^2d^2$ and hence $d = n/2$ because of $|E| = n^2/4$. But this immediately implies $G = K_{n/2, n/2}$.

$k = 3$: **Second proof** (Folklore). Let A be a largest independent set, $|A| = \alpha$. Since G is triangle-free, we have $d_i \leq \alpha$ for all i . The set $B = V \setminus A$ meets every edge of G , whence we obtain $|E| \leq \sum_{i \in B} d_i$ by counting in two ways. Setting $|B| = \beta = n - \alpha$ we obtain by the inequality of the arithmetic-geometric mean

$$|E| \leq \sum_{i \in B} d_i \leq \alpha \cdot \beta \leq \left(\frac{\alpha + \beta}{2} \right)^2 = \frac{n^2}{4}. \quad \square$$

Now we turn to the proofs of the general case (1).

First proof (Turán 1941). We use induction on n . (1) is trivially true for small n . Let G be a graph on $V = \{1, \dots, n\}$ without k -cliques with a maximal number of edges. G certainly contains $(k-1)$ -cliques, since otherwise we could add edges. Let A be a $(k-1)$ -clique, $B = V \setminus A$, $|B| = n - k + 1$ (Figure 3).

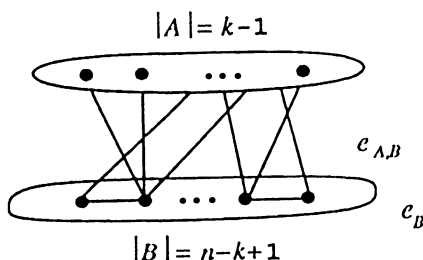


Figure 3.

A contains $\binom{k-1}{2}$ edges, and we now estimate the edge-number e_B in B and the edge-number $e_{A,B}$ between A and B . By induction, we have $e_B \leq (k-2)/2(k-1)(n-k+1)^2$. Since G has no k -clique, every $j \in B$ is adjacent to at most $k-2$ vertices in A , and we obtain $e_{A,B} \leq (k-2)(n-k+1)$. Altogether, this yields

$$|E| \leq \binom{k-1}{2} + \frac{k-2}{2(k-1)}(n-k+1)^2 + (k-2)(n-k+1), \quad (6)$$

which is precisely $(k-2)/2(k-1)n^2$. \square

Second proof (Erdős 1970). This proof makes use of the structure of the Turán graphs. Let $m \in V$ with $d_m = \max_{1 \leq j \leq n} d_j$. We denote by S the neighbors of m , $|S| = d_m$, and set $T = V \setminus S$. As G contains no k -clique, and m is adjacent to all of S , we note that S contains no $(k-1)$ -clique. We now construct the following graph H on V (see Figure 4). H corresponds to G on S and contains all edges between S and T , but no edges within T .

In other words, T is an independent set in H , and we conclude that H has again no k -cliques. Let d'_j be the degree of j in H . If $j \in S$, then we certainly have $d'_j \geq d_j$ by the construction of H , and for $j \in T$, we see $d'_j = |S| = d_m \geq d_j$ by the choice of m . We infer $|E(H)| \geq |E|$, and conclude that among all graphs with a maximal number of edges, there must be one of the form of H . Applying induction on S , we thus infer that among the graphs with a maximal number of edges there is a graph $K_{n_1, \dots, n_{k-1}}$, which implies $|E| \leq \sum_{i \neq j} n_i n_j$ and therefore (1). \square

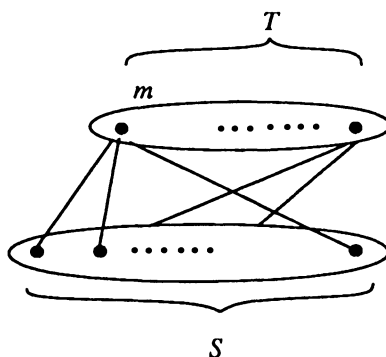


Figure 4.

We note that this proof yields the full statement $|E| \leq |E(H)|$, $H =$ Turán graph.

Third proof (Moon-Moser 1962). This proof generalizes the idea of the first proof for $k = 3$ and yields a quantitative estimate for the number of h -cliques. Let G be any graph on $V = \{1, \dots, n\}$ and denote by \mathcal{E}_h the set of h -cliques in G with $|\mathcal{E}_h| = C_h$. As examples we have $C_1 = n$, $C_2 = |E|$, $C_3 =$ number of triangles. For $A \in \mathcal{E}_h$ we denote by $d(A)$ the number of $(h + 1)$ -cliques containing A . Counting in two ways we obtain

$$\sum_{A \in \mathcal{E}_h} d(A) = (h + 1)C_{h+1} \quad (h \geq 1), \quad (7)$$

in generalization of (2). For $A \in \mathcal{E}_h$ ($h \geq 2$) let us denote by $A^{(1)}, \dots, A^{(h)}$ the $(h - 1)$ -cliques contained in A .

Claim. For any graph G

$$\frac{C_{h+1}}{C_h} \geq \frac{1}{h^2 - 1} \left(h^2 \frac{C_h}{C_{h-1}} - n \right) \quad (h \geq 2). \quad (8)$$

Consider $A \in \mathcal{E}_h$, $B = V \setminus A$, $|B| = n - h$. Among the vertices $j \in B$ there are precisely $d(A)$ vertices which are adjacent to all of A . Every other vertex in B is adjacent to at most one $(h - 1)$ -clique $A^{(i)}$, thereby forming an h -clique (Figure 5). We thus obtain (note $- 1$ because of $A^{(i)} \subseteq A$)

$$\sum_{i=1}^h (d(A^{(i)}) - 1 - d(A)) + d(A) \leq n - h,$$

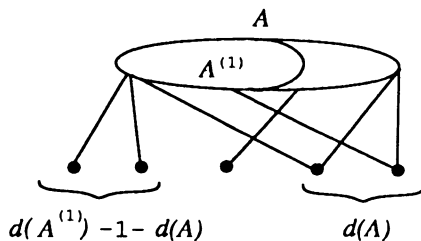


Figure 5.

hence

$$\sum_{i=1}^h d(A^{(i)}) - (h-1)d(A) \leq n.$$

Summation over $A \in \mathcal{C}^{(h)}$ yields

$$\sum_{A \in \mathcal{C}_h} \sum_{i=1}^h d(A^{(i)}) - (h-1) \sum_{A \in \mathcal{C}_h} d(A) \leq nC_h. \quad (9)$$

As in (4) we conclude

$$\sum_{A \in \mathcal{C}_h} \sum_{i=1}^h d(A^{(i)}) = \sum_{B \in \mathcal{C}_{h-1}} d(B)^2, \quad (10)$$

and by (7) we have

$$(h-1) \sum_{A \in \mathcal{C}_h} d(A) = (h^2-1)C_{h+1}. \quad (11)$$

Substituting (10) and (11) into (9) gives us

$$\sum_{B \in \mathcal{C}_{h-1}} d(B)^2 \leq nC_h + (h^2-1)C_{h+1}. \quad (12)$$

By the Cauchy-Schwarz inequality applied to the vectors $(d(B))$, (1) of length C_{h-1} , we finally obtain

$$nC_h + (h^2-1)C_{h+1} \geq \sum_{B \in \mathcal{C}_{h-1}} d(B)^2 \geq \frac{1}{C_{h-1}} \left(\sum_{B \in \mathcal{C}_{h-1}} d(B) \right)^2 = \frac{h^2 C_h^2}{C_{h-1}},$$

which is precisely (8).

In order to prove (1) we must find a relationship between (8) and the edge-number $|E|$. Let us set

$$|E| = \left(1 - \frac{1}{\vartheta}\right) \frac{n^2}{2} (\vartheta \in \mathbb{R}). \quad (13)$$

Since the right-hand side of (13) is increasing in ϑ , we must thus prove $\vartheta \leq k-1$ for graphs without k -cliques.

Claim. *We have*

$$\frac{C_{h+1}}{C_h} \geq \frac{\vartheta-h}{\vartheta} \frac{n}{h+1} (h \geq 1). \quad (14)$$

For $n=1$ we have $C_2 = |E|$, $C_1 = n$, and (14) is satisfied with equality by the definition of ϑ . Using (8) and induction on h we infer

$$\begin{aligned} \frac{C_{h+1}}{C_h} &\geq \frac{1}{h^2-1} \left(h^2 \frac{\vartheta-h+1}{\vartheta} \frac{n}{h} - n \right) = \frac{1}{h^2-1} \frac{(\vartheta-h)(h-1)n}{\vartheta} \\ &= \frac{\vartheta-h}{\vartheta} \cdot \frac{n}{h+1}, \end{aligned}$$

as claimed.

Now, if G contains no k -clique, then $C_k = 0$, and we infer $\vartheta \leq k-1$ from (14) for $h+1=k$. \square

EXAMPLE. Consider (8) for $h = 2$. In this case, the inequality states that any graph satisfies

$$C_3 \geq \frac{|E|}{3} \left(\frac{4|E|}{n} - n \right) = \frac{|E|}{3n} (4|E| - n^2).$$

We conclude that a graph G on an even number n of vertices with $|E| = n^2/4 + 1$ not only contains one triangle (as it must by Turán's Theorem), but more than $n/3$. If we add one edge to $K_{n/2, n/2}$, then we obtain $n/2$ triangles, and it can be easily shown that this holds for any graph with $n^2/4 + 1$ edges.

So far, the proofs have employed counting techniques, the following three proofs use entirely different ideas.

Fourth proof (Motzkin-Straus 1965). Let G be an arbitrary graph on $V = \{1, \dots, n\}$. By $\omega = \omega(G)$ we denote the number of vertices in a largest clique of G , $\omega(G)$ is called the *clique-number*. Now, we associate to each $i \in V$ a variable x_i (over \mathbb{R}) and consider the function $f(x_1, \dots, x_n) = 2\sum_{ij \in E} x_i x_j$.

Claim. We have

$$1 - \frac{1}{\omega} = \max \left(2 \sum_{ij \in E} x_i x_j : \sum_{i=1}^n x_i = 1, x_i \geq 0 \text{ for all } i \right). \quad (15)$$

Since f is continuous on a compact set, there exists x with $f(x) = \max$. Among all such vectors x , we choose one with a maximal number of $x_i = 0$. Let $C = \{i \in V : x_i > 0\}$. We show first that C is a clique. Suppose this is false with $1, 2 \in C$ but $12 \notin E$. For any $t \in \mathbb{R}$ in the range $-x_1 \leq t \leq x_2$ the vector $x_t = (x_1 + t, x_2 - t, x_3, \dots, x_n)$ satisfies the conditions in (15), and furthermore, $f(x_t)$ is a linear function in t , since the product $(x_1 + t)(x_2 - t)$ does not appear in $f(x_t)$ because of $12 \notin E$. Since by the choice of x , $f(x_t)$ assumes the maximum at $t = 0$ (i.e. in the interior) we conclude that $f(x_t)$ is, in fact, constant for all t . For $t = x_2$, $\bar{x} = (x_1 + x_2, 0, x_3, \dots, x_n)$, we therefore obtain $f(\bar{x}) = f(x)$, contradicting the choice of x .

We can thus assume $f(x) = \max$ with $C = \{i : x_i > 0\}$ a clique. Since

$$1 = (x_1 + \dots + x_n)^2 = 2 \sum_{ij \in C} x_i x_j + \sum_{i \in C} x_i^2$$

we conclude that $f(x)$ is maximal if and only if $\sum_{i \in C} x_i^2$ is minimal. Under the assumption $\sum_{i \in C} x_i = 1$ this is clearly the case for $x_i = 1/|C|$, and we obtain

$$f(x) = 1 - \sum_{i \in C} x_i^2 = 1 - \frac{1}{|C|} \leq 1 - \frac{1}{\omega}$$

with equality for $|C| = \omega$, which is what we wanted to prove.

Inequality (1) is now an immediate consequence. Setting $x_i = 1/n$, we have $f(x) = 2|E|/n^2$ and therefore

$$\frac{2|E|}{n^2} = f(x) \leq 1 - \frac{1}{k-1} = \frac{k-2}{k-1},$$

since G contains no k -clique. □

Fifth proof (Li-Li 1981, Kleitman-Lovász 1994). The basis for this proof is again an algebraic structure. To every vertex $i \in V$ of the graph G we again associate a

variable x_i and consider the polynomial

$$p_G(x_1, \dots, x_n) = \sum_{i < j, ij \notin E} (x_i - x_j). \quad (16)$$

The fundamental observation on the polynomial p_G is the following obvious fact:

$$\omega(G) \leq k - 1 \Leftrightarrow \text{the identification } x_{i_1} = \dots = x_{i_k} \text{ of any } k \text{ variables in } p_G \text{ yields the zero-polynomial.} \quad (17)$$

Let $P(n, k)$ be the set of real polynomials in n variables which satisfy the right-hand side of (17). $P(n, k)$ is clearly an ideal in $\mathbb{R}[x_1, \dots, x_n]$. Let $\mathcal{H}(n, k)$ be the following family of graphs on $V = \{1, \dots, n\}$: H is in $\mathcal{H}(n, k)$ if and only if the vertex-set V can be partitioned into $k - 1$ disjoint independent subsets (in the language of graph theory, this means H is $(k - 1)$ -partite or $(k - 1)$ -colorable). In particular, all our graphs $K_{n_1, \dots, n_{k-1}}$ are in \mathcal{H} and therefore all Turán graphs. By our remarks on the graphs $K_{n_1, \dots, n_{k-1}}$ we can therefore state

$$|E(H)| \leq \frac{(k - 2)n^2}{2(k - 1)} \text{ for all } H \in \mathcal{H}(n, k). \quad (18)$$

By $\hat{P}(n, k)$ we denote the ideal in $\mathbb{R}[x_1, \dots, x_n]$ generated by $\{p_H : H \in \mathcal{H}(n, k)\}$. Since we have $\omega(H) \leq k - 1$ for any such graph, we infer $\hat{P}(n, k) \subseteq P(n, k)$.

Claim. We have $P(n, k) = \hat{P}(n, k)$.

Before proving this claim, let us see how Turán's theorem follows from it. Let G be a graph with $\omega(G) \leq k - 1$. Then $p_G \in P(n, k) = \hat{P}(n, k)$, i.e.

$$p_G = \sum_{i=1}^m q_i p_{H_i} \text{ with } H_i \in \mathcal{H}(n, k), q_i \in \mathbb{R}[x_1, \dots, x_n]. \quad (19)$$

By (16), p_G is a homogeneous polynomial of degree $(p_G) = \binom{n}{2} - |E(G)|$, and analogously degree $(p_{H_i}) = \binom{n}{2} - |E(H_i)|$. We thus infer from (19), $\binom{n}{2} - |E(G)| \geq \binom{n}{2} - |E(H_i)|$ for some i , and therefore (1) from (18).

Let $f \in P(n, k)$. To prove $f \in \hat{P}(n, k)$ we use induction on n . For $n = 2$ there is nothing to prove. For a subset $S \subseteq \{1, \dots, n - 1\}$ we denote by f_S the polynomial which results from f by identifying $x_n = x_i$ for all $i \in S$. Clearly, $f_S \in P(n, k)$ and hence $f_S \in \hat{P}(n, k)$ for $S \neq \emptyset$ by induction (note $\hat{P}(n - 1, k) \subseteq \hat{P}(n, k)$). Now consider the polynomial

$$g = \sum_{S \subseteq \{1, \dots, n-1\}} (-1)^{|S|} f_S. \quad (20)$$

Cancelling terms we see that every identification $x_n = x_i$ ($i = 1, \dots, n - 1$) in g yields the zero-polynomial. We conclude that $(x_1 - x_n) \dots (x_{n-1} - x_n)$ divides g , hence

$$g = (x_1 - x_n) \dots (x_{n-1} - x_n) h. \quad (21)$$

Since $f_S \in P(n, k)$ for all S , we have $g \in P(n, k)$ by (20), whence h becomes by (21) the zero-polynomial whenever we identify k of the variables x_1, \dots, x_{n-1} in h . Expanding h with respect to x_n , we see that every coefficient polynomial p of a power x_n^l lies in $P(n - 1, k)$ and hence in $\hat{P}(n - 1, k)$ by induction. We conclude that the polynomial g is a sum of expressions

$$q(x_1 - x_n) \dots (x_{n-1} - x_n) p_{\bar{H}}, \quad (22)$$

with $\bar{H} \in \mathcal{H}(n - 1, k)$, $q \in \mathbb{R}[x_1, \dots, x_n]$.

Adding the vertex n to each such \bar{H} without edges from n to \bar{H} , we obtain $(x_1 - x_n) \dots (x_{n-1} - x_n) p_{\bar{H}} = p_H$ with $H \in \mathcal{H}(n, k)$. This now implies $g \in \hat{P}(n, k)$ by (22), and thus

$$f = g - \sum_{S \neq \emptyset} (-1)^{|S|} f_S \in \hat{P}(n, k),$$

as claimed. \square

REMARK. We note that this proof again yields the full implication of (1), that the Turán graphs attain the maximal number of edges, and it can be shown that the polynomials p_H , $H = \text{Turán graph}$, already generate the ideal $P(n, k)$.

Sixth proof (Alon-Spencer 1992). Our last and perhaps most elegant proof uses ideas from probability theory. Let G be an arbitrary graph on $V = \{1, \dots, n\}$.

Claim. *We have*

$$\omega(G) \geq \sum_{i=1}^n \frac{1}{n - d_i}. \quad (23)$$

We choose with equal probability $1/n!$ a permutation $\pi_1, \pi_2, \dots, \pi_n$ of V and construct the following set C . We put π_i into C if and only if π_i is adjacent to all π_j ($j < i$). By definition C is a clique in G . Let $X = |C|$ be the corresponding random variable. We have $X = \sum_{i=1}^n X_i$, where X_i is the indicator random variable of i , i.e. $X_i = 1$ or 0 depending on $i \in C$ or $i \notin C$. Now we note $i \in C$ with respect to the permutation (π_1, \dots, π_n) iff i appears *before* all $n - 1 - d_i$ non-neighbors of i , or in other words, if i is the *first* among i and its non-neighbors. We conclude $EX_i = 1/n - d_i$ for the expectation and hence

$$E(|C|) = EX = \sum_{i=1}^n EX_i = \sum_{i=1}^n \frac{1}{n - d_i}$$

by the linearity of expectation. Consequently, there must be a clique C with at least $E(|C|)$ vertices, and this is just our claim (23).

To deduce Turán's theorem from (23) we use the Cauchy-Schwarz inequality in the form

$$n^2 = \left(\sum \sqrt{x_i} \sqrt{x_i^{-1}} \right)^2 \leq \sum x_i \cdot \sum x_i^{-1}$$

with $x_i = n - d_i$. Indeed, (23) and (2) imply

$$\omega(G) \geq \frac{n^2}{\sum_{i=1}^n n - d_i} = \frac{n^2}{n^2 - 2|E|}. \quad (24)$$

If G has no k -clique, then $\omega(G) \leq k - 1$ and (24) reduces precisely to (1). \square

REMARK. Inequality (23) was first proved in Wei [10] by successively removing vertices similar to the second proof.

REFERENCES

1. N. Alon-J. Spencer: *The Probabilistic Method*. Wiley-Interscience 1992.
2. B. Bollobás: *Extremal graph Theory*. Academic Press 1978.
3. P. Erdős: On the graph theorem of Turán (in Hungarian). *Math. Fiz. Lapok* 21 (1970), 249–251.

4. S. R. Li-W. W. Li: Independence number of graphs and generators of ideals. *Combinatorica* 1 (1981), 55–61.
5. L. Lovász: Stable sets and polynomials. *Discrete Math.* 124 (1994), 137–153.
6. W. Mantel: Problem 28. *Wiskundige Opgaven* 10 (1906), 60–61.
7. J. W. Moon-L. Moser: On a problem of Turán. *Publ. Math. Inst. Hungar. Acad. Sci.* 7 (1962), 283–286.
8. T. S. Motzkin-E. G. Straus: Maxima for graphs and a new proof of a theorem of Turán. *Canad. J. Math.* 17 (1965), 533–540.
9. P. Turán: On an extremal problem in graph theory (in Hungarian). *Math. Fiz. Lapok* 48 (1941), 436–452.
10. V. K. Wei: A lower bound on the stability number of a simple graph. *Bell Lab. Tech. Mem.* 81-11217-9 (1981).

II. Mathematisches Institut
Freie Universität Berlin
Arnimallee 3 D-14195
Berlin, GERMANY
aigner@math.fu-berlin.de

The fact is, although DNA testing may be as foolproof as fingerprinting, it doesn't cause excitement. It's difficult to respond to. It's like advanced math, brilliant but boring, astonishing but passionless. It made everyone eager to move on to the next phase of the trial, which consisted of autopsy pictures . . .

From "If the Gloves Fit" by Dominick Dunne, in *Vanity Fair*/August 1995.

**Submitted by J. Foster
 Weber State University**

**Answer to Picture Puzzle
 (p. 797)**

A. S. Bessicovitch.

NOTES

Edited by: John Duncan

More on Kummer's Test

Hans Samelson

These are some remarks related to the interesting Note "*Kummer's Test Gives Characterizations for Convergence and Divergence of all Positive Series*" by Jengching Tong ([4]).

(All sequences below have positive terms.) In slightly changed notation (see below), Kummer's test for convergence (developed in [2]) says: a series $\sum_1 a_k$ is convergent if there is a (strongly decreasing) sequence $\{c_k\}_0^\infty$ with $c_{k-1} - c_k \geq c \cdot a_k$; the positive constant c is not important, and we shall mostly take it equal to 1. For a proof note that $\sum(c_{k-1} - c_k)$ is convergent (its partial sums are bounded by c_0 , by telescoping), and then, by the inequality, $\sum a_k$ also converges (comparison test). (This is essentially the proof given by U. Dini [1], p. 66 (Opere) and by O. Stolz [3], p. 259. Kummer assumed $\lim c_k = 0$; however one can always adjust the c_k so that this holds, as noted by Dini, [1], p. 49 (Opere).)

But note now that any convergent series $\sum b_k$ can be written in the form $\sum(c_{k-1} - c_k)$ as above; take for c_k the series remainder $\sum_{k+1}^\infty a_i$. Thus Kummer's convergence criterion can be stated as: a series $\sum a_k$ converges if (and trivially only if) there is a converging series $\sum b_k$ that majorizes it term by term; it is exactly the basic comparison test (which had been introduced explicitly by Cauchy in 1821)! (Kummer's own proof does not quote the comparison test; he proves his assertion by estimating the series sections $\sum_{m+1}^n a_k$; his argument is actually quite similar to that of Dini and Stolz.)

The opening paragraph of Kummer's paper (written when he was 23, two years after his Ph.D., while he was teaching at a Gymnasium in a small town in Silesia) is quite striking; he says that since no universally valid criterion for convergence or divergence has ever been found he was looking for a method to test an arbitrary series, the test being contained in the following theorems. (Indeed he says later on that his convergence test and his divergence test [see below] give a decision for every series.) I wonder whether he realized how close to the comparison test he was.

His real contribution and insight here was to take the b_k in the form $c_{k-1} - c_k$ and to write the c_k as products $m_k a_k$ (actually he wrote his condition as $f(k) = m_k a_k / a_{k+1} - m_{k+1} \geq c$, or rather " $f(k) > 0$ for $k = \infty$ "!, assuming apparently that the sequence $f(k)$ must converge); as he himself, Dini, and others showed, fairly simple choices of the m_k lead to many good tests for specific series or classes of series. (Taking $m_k = 1$, one gets Cauchy's ratio criterion, comparison with the geometric series; with $m_k = k$ one is led to Raabe's criterion, etc (see [4]).)

Kummer also gave a test for divergence (not the one quoted in [4]; that one was introduced by Dini in [1]). It seems to me that the proof contains a mistake that makes it invalid; Kummer seems to assume that the sequence $m_k a_k / a_{k+1} - m_{k+1}$

is monotone decreasing, for which behavior I can't find any reason. Dini, who recapitulates this proof in [1], seems to assume that for a sequence h_k that goes to 0 the quotients h_m/h_n for $m > n$ are bounded.

And as a last remark: Dini's test for divergence says that $\sum a_k$ is divergent if there exists a (weakly) monotone increasing sequence $c_k = m_k a_k$ with $\sum 1/m_k$ ($= \sum a_k/c_k$) diverging. The monotonicity condition can be written as $a_k/a_{k+1} \geq (1/m_k)/(1/m_{k+1})$, and by a standard argument that makes $\sum a_k$ diverge also. (That is Dini's proof.)

Let us here replace the monotonicity of the c_k by the weaker condition that the c_k are bounded below by a positive number, say by 1. Then we have $a_k \geq a_k/c_k$, with the series of the latter terms diverging by assumption, and thus a slightly improved version of Dini's criterion also turns out to be identical with the comparison test (divergence version)! The interesting aspect is again that writing the c_k as $m_k a_k$ makes it easy to set up specific tests.

REFERENCES

1. U. Dini, *Sulle Serie a Termini Positivi*, Annali d. Università Toscana vol. 9, 1867, 41–76 = Opere, Ed. Cremonese, Rome 1953, vol. I, 29–69.
2. E. E. Kummer, *Über die Convergenz und Divergenz der unendlichen Reihen*, Journ. für die reine und angewandte Mathematik vol. 13 (1835), 171–84 = Collected Papers, Springer Verlag New York 1975, vol. II, 47–60.
3. O. Stolz, *Vorlesungen über allgemeine Arithmetik* vol. I, Teubner, Leipzig 1885.
4. Jingcheng Tong, *Kummer's Test Gives Characterizations for Convergence or Divergence of all Series*, The American Mathematical Monthly vol. 101, 1994, 450–452.

Department of Mathematics
Stanford University
Stanford, CA 94305
samelson@gauss.stanford.edu

The Derivation of the Exponential Map of Matrices

G. M. Tuynman

The exponential map, which links the Lie algebra to its Lie group, is of course an analytic map, and as such it has a derivative. Although the explicit expression for this derivative is not so complicated, the way to obtain it seems long and difficult. For instance, in [H] affine connections and differential equations are used, in [P] a Taylor expansion is used and terms of order 2 ($\mathcal{O}(t^2)$) are neglected, in [V] a complicated analysis of Taylor series using enveloping algebras is used; in [MT] a rather simple argument using differential equations is used, but this argument is only valid for matrices.

We present here a rather elementary way to obtain this derivative. For ease of exposition we will do it for matrices, but only cosmetic changes are needed to make it a valid computation for any Lie group.

Theorem. Let $\exp \equiv e: M(n, \mathbf{R}) \rightarrow M(n, \mathbf{R})$ denote the exponential map on $n \times n$ matrices with real entries. Then:

$$\left. \frac{d}{dt} \right|_{t=0} e^{X+tY} = e^X \cdot \left(\frac{1 - e^{-\text{ad}(X)}}{\text{ad}(X)} \right)(Y),$$

where $\text{ad}(X)$ denotes the adjoint representation $Y \mapsto \text{ad}(X)(Y) \equiv X \cdot Y - Y \cdot X$, and where the quotient should be interpreted as the formal power series.

Proof: We introduce the matrix $\Delta(X, Y)$ defined as:

$$\Delta(X, Y) = e^{-X} \cdot \left. \frac{d}{dt} \right|_{t=0} e^{X+tY}.$$

The map Δ is obviously continuous in X and Y , and, moreover, it is linear in Y by definition of the derivative. Applying the Leibnitz rule to the equality $e^{X+tY} = \exp((1/n)X + t(1/n)Y)^n$, valid for any $n \in \mathbf{Z}$, gives us:

$$\left. \frac{d}{dt} \right|_{t=0} e^{X+tY} = \sum_{k=0}^{n-1} \exp\left(\frac{1}{n}X\right)^{n-1-k} \cdot \left(\left. \frac{d}{dt} \right|_{t=0} \exp\left(\frac{1}{n}X + t\frac{1}{n}Y\right) \right) \cdot \exp\left(\frac{1}{n}X\right)^k.$$

Using the definition of Δ we then compute:

$$\begin{aligned} e^{-X} \cdot \left. \frac{d}{dt} \right|_{t=0} e^{X+tY} &= \sum_{k=0}^{n-1} \exp\left(\frac{1}{n}X\right)^{-k} \cdot \Delta\left(\frac{1}{n}X, \frac{1}{n}Y\right) \cdot \exp\left(\frac{1}{n}X\right)^k \\ &= \frac{1}{n} \cdot \sum_{k=0}^{n-1} \text{Ad}(e^{-X/n})^k \left(\Delta\left(\frac{1}{n}X, Y\right) \right) \\ &= \left(\frac{1 - \text{Ad}(e^{-X})}{n(1 - \text{Ad}(e^{-X/n}))} \right) \left(\Delta\left(\frac{1}{n}X, Y\right) \right) \\ &\xrightarrow{n \rightarrow \infty} \left(\frac{1 - e^{-\text{ad}(X)}}{\text{ad}(X)} \right) (\Delta(0, Y)). \end{aligned}$$

To obtain the second equality we used the linearity of Δ in Y and the definition of the Adjoint representation: $\text{Ad}(B)(A) = B \cdot A \cdot B^{-1}$. For the third equality we used the formula for the sum of a geometric progression with factor $\text{Ad}(e^{-X/n})$. For the limit we used the continuity of Δ in X , the fact that ad is the derivative of Ad (for the limit $n \rightarrow \infty$ in the denominator!), and that the exponential map intertwines ad and Ad : $\text{Ad}(e^X) = e^{\text{ad}(X)}$. Since an elementary calculation shows that $\Delta(0, Y) = Y$, the theorem follows when we multiply by e^X .

Remark. Readers who feel uneasy in taking the limit $n \rightarrow \infty$ need only check the following convergence of power series of a single complex variable z :

$$\frac{e^z - 1}{n(e^{z/n} - 1)} = \sum_{k=0}^{n-1} \frac{1}{n} (e^{z/n})^k = \sum_{i=0}^{\infty} \left(\sum_{k=0}^{n-1} \frac{k^i}{i! n^{i+1}} \right) \cdot z^i \xrightarrow{n \rightarrow \infty} \sum_{i=0}^{\infty} \frac{z^i}{(i+1)!} = \frac{e^z - 1}{z}.$$

REFERENCES

- [H] S. Helgason, *Differential geometry, Lie groups, and symmetric spaces*, Academic Press, Orlando, 1978.
 [MT] R. Mneimné & F. Testard, *Introduction à la théorie des groupes de Lie classiques*, Hermann, Paris, 1986.

- [P] M. Postnikov, *Leçons de géométrie: Groupes et algèbres de Lie*, Editions MIR, Moscou, 1982, 1985.
- [V] V. S. Varadarajan, *Lie groups, Lie algebras, and their representations*, Prentice-Hall, Englewood Cliffs, 1974; Reprinted as GTM volume 102, Springer Verlag, Berlin.

URA D0751 au CNRS & UFR de Mathématiques
 Université de Lille I
 F-59655 Villeneuve d'Ascq Cedex
 France
 gmt@gat.univ-lille1.fr

The Kantorovich Inequality

Vlastimil Pták

The inequality appears first in a survey article on functional analysis and applied mathematics by L. V. Kantorovič; it is used in investigations concerning the condition number of operators and has important applications in estimating convergence of methods of steepest descent for solving equations. In a number of subsequent papers the connection of the inequality with an inequality given by Pólya and Szegő was cleared up and a number of proofs, some of considerable complexity, of the inequality and of different variants thereof appeared in the literature.

In view of the importance of the inequality one more note on the subject might be of interest. It is not difficult to see that the result is essentially based on the inequality between the geometric and arithmetic mean; to emphasise this we restate it in a form using the two means which immediately suggests a simple and natural proof.

The Kantorovich inequality. Suppose $x_1 < x_2 < \cdots < x_n$ are given positive numbers. Let $\lambda_1, \dots, \lambda_n \geq 0$ and $\sum \lambda_j = 1$. Then

$$\left(\sum \lambda_j x_j \right) \left(\sum \lambda_j x_j^{-1} \right) \leq A^2 G^{-2}$$

where $A = \frac{1}{2}(x_1 + x_n)$ and $G = (x_1 x_n)^{1/2}$.

Proof: Observe that the inequality is homogeneous in the sense that it is invariant with respect to replacing each x_j by a positive multiple αx_j . Accordingly it is possible to assume that $G = 1$ so that $x_n = 1/x_1$. Each x between x_1 and $1/x_1$ satisfies

$$x + \frac{1}{x} \leq x_1 + \frac{1}{x_1}.$$

It follows that $\sum \lambda_j x_j + \sum \lambda_j x_j^{-1} \leq x_1 + 1/x_1 = 2A$. The conclusion follows by an application of the geometric—arithmetic mean inequality.

REFERENCES

1. F. L. Bauer: A further generalization of the Kantorovich inequality, *Numer. Math.* 3 (1961), 117–119.
2. F. L. Bauer and A. S. Householder: Some inequalities involving the euclidean condition of a matrix, *Numer. Math.* 2 (1960), 308–311.
3. W. Greub and W. Reinboldt: On a generalization of an inequality of L. V. Kantorovich, *Proc. Amer. Math. Soc.* 10 (1959), 407–413.
4. P. Henrici: Two remarks on the Kantorovich inequality, *Amer. Math. Monthly*, 68 (1961), 904–906.
5. A. S. Householder and F. L. Bauer: On certain iterative methods for solving linear systems, *Numer. Math.* 2 (1960), 55–59.
6. L. V. Kantorovič: Functional analysis and applied mathematics (in Russian), *Uspechi mat. nauk*, 3 (1948), 89–185.
7. M. Newman: Kantorovich's inequality, *J. Res. Natl. Bur. Standards*, 64B (1960), 33–34.
8. G. Pólya and G. Szegő: *Aufgaben und Lehrsätze der Analysis*, Berlin, 1925.
9. A. H. Schopf: On the Kantorovich inequality, *Numer. Math.* 2 (1960), 344–346.
10. W. G. Strang: On the Kantorovich inequality, *Proc. Amer. Math. Soc.* 11 (1960), 468.

Academy of Sciences of the Czech Republic
Institute of Mathematics
Žitná 25
11567 Praha 1
The Czech Republic

On The Generalized Inverse Form of the Equations of Constrained Motion

Robert Kalaba and Rong Xu

1. INTRODUCTION. It has often been observed that seemingly abstract concepts and principles prove to be of paramount importance in practical applications. The close ties between constrained motion and generalized inverses of matrices may be a case in point.

In 1829 C. F. Gauss formulated his celebrated principle of least constraint for handling static and dynamic problems for constrained mechanical systems. The principle takes the form of a minimization problem. The seemingly abstract notion of the generalized inverse of a matrix proves to be crucial in dealing with Gauss' principle and in understanding the complex interactions between applied and constraint forces.

In this note, we present Gauss' principle and then indicate the role of generalized inverses in its further development.

2. GAUSS' PRINCIPLE AND GENERALIZED INVERSES OF MATRICES [1]. Consider a system of p particles. Let the mass of the i^{th} particle be m_i and the external force acting on it be f_i . We use Cartesian coordinates. If there were no

constraint on the particles, the acceleration of the i^{th} particle would be $\mathbf{a}_i = \mathbf{f}_i/m_i$. In matrix form this is

$$\mathbf{F} = \mathbf{M}\mathbf{a}, \quad (1)$$

where, with $n = 3p$,

$$\mathbf{F}_{n \times 1} = \begin{bmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \vdots \\ \mathbf{f}_p \end{bmatrix}, \quad (2)$$

$$\mathbf{M}_{n \times n} = \begin{bmatrix} m_1 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & m_1 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & m_1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & m_2 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & m_2 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & m_2 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & m_p & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & m_p & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & m_p \end{bmatrix}, \quad (3)$$

and

$$\mathbf{a}_{n \times 1} = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_p \end{bmatrix}. \quad (4)$$

However, if the particles are subject to constraints, a task of analytical mechanics is to determine the actual acceleration of the i^{th} particle, $\ddot{\mathbf{x}}_i$, at any time t . Gauss' principle of least constraint [2] states that the actual accelerations $\ddot{\mathbf{x}}_1, \ddot{\mathbf{x}}_2, \dots, \ddot{\mathbf{x}}_p$ minimize G where

$$G = \sum_{i=1}^p m_i (\ddot{\mathbf{x}}_i - \mathbf{a}_i)^T (\ddot{\mathbf{x}}_i - \mathbf{a}_i), \quad (5)$$

subject to whatever the constraints might be on the accelerations. As usual, superscript T denotes transposition. Eq. (5) is identical to

$$G = (\ddot{\mathbf{x}} - \mathbf{a})^T \mathbf{M}(\ddot{\mathbf{x}} - \mathbf{a}), \quad (6)$$

where

$$\ddot{\mathbf{x}}_{n \times 1} = \begin{bmatrix} \ddot{\mathbf{x}}_1 \\ \ddot{\mathbf{x}}_2 \\ \vdots \\ \ddot{\mathbf{x}}_p \end{bmatrix}. \quad (7)$$

In Lagrangian mechanics, the constraints take the form $\varphi_k(t, x, \dot{x}) = 0$, where $k = 1, 2, \dots, m$, in which m is the number of constraints. Through differentiation

with respect to t , this leads to the constraints on the acceleration, taking the form of a consistent linear algebraic system

$$\mathbf{A}\ddot{\mathbf{x}} = \mathbf{b}, \quad (8)$$

where \mathbf{A} is an $m \times n$ matrix and \mathbf{b} is an $m \times 1$ vector. Both \mathbf{A} and \mathbf{b} may depend upon the time t , the particles' displacement vector \mathbf{x} and the velocity vector $\dot{\mathbf{x}}$; and the matrix \mathbf{A} need not be of full rank. Given initial conditions on \mathbf{x} and $\dot{\mathbf{x}}$, use of the differentiated form of the constraints does not cause any loss in generality. In these cases, Gauss' principle takes the form of minimizing G subject to the linear constraints (8).

To recast Gauss' principle, let

$$\mathbf{y} = \mathbf{M}^{1/2}(\ddot{\mathbf{x}} - \mathbf{a}), \quad (9)$$

so that

$$\ddot{\mathbf{x}} = \mathbf{M}^{-1/2}\mathbf{y} + \mathbf{a}. \quad (10)$$

Consequently, Eqs. (6) and (8) are equivalent to

$$G = \mathbf{y}^T \mathbf{y}, \quad (11)$$

and

$$\mathbf{A}\mathbf{M}^{-1/2}\mathbf{y} = \mathbf{b} - \mathbf{A}\mathbf{a}. \quad (12)$$

Gauss' principle is then reduced to the problem of finding the shortest length vector \mathbf{y} such that the consistent linear Eq. (12) is satisfied. As is known [3], the solution to this problem can be expressed by

$$\mathbf{y} = (\mathbf{A}\mathbf{M}^{-1/2})^+ (\mathbf{b} - \mathbf{A}\mathbf{a}), \quad (13)$$

where $(\mathbf{A}\mathbf{M}^{-1/2})^+$ is the Moore-Penrose generalized inverse* of the matrix $\mathbf{A}\mathbf{M}^{-1/2}$. This is the only property of the generalized inverse that is used! Substituting Eq. (9) into Eq. (13) and rearranging it, we obtain the actual acceleration vector $\ddot{\mathbf{x}}$ in the form

$$\ddot{\mathbf{x}} = \mathbf{a} + \mathbf{M}^{-1/2}(\mathbf{A}\mathbf{M}^{-1/2})^+ (\mathbf{b} - \mathbf{A}\mathbf{a}). \quad (14)$$

3. AN APPLICATION. Let us use a simple example to illustrate the applicability and utility of the formula given by Eq. (14). Suppose we would like to determine the equations of motion of a spherical pendulum with length l and mass m . The rectangular coordinates of the mass point are specified by (x_1, x_2, x_3) . The following is a general procedure.

First, we identify the mass matrix and the free motion acceleration vector. In this case, they are

$$\mathbf{M}_{3 \times 3} = \begin{bmatrix} m & 0 & 0 \\ 0 & m & 0 \\ 0 & 0 & m \end{bmatrix}, \quad (15)$$

*Given an $m \times n$ matrix \mathbf{B} , there always exists a unique $n \times m$ matrix \mathbf{B}^+ , called the Moore-Penrose generalized inverse of \mathbf{B} , which satisfies the following four basic properties:

$$\mathbf{B}\mathbf{B}^+\mathbf{B} = \mathbf{B}; \quad \mathbf{B}^+\mathbf{B}\mathbf{B}^+ = \mathbf{B}^+ \quad (\mathbf{B}\mathbf{B}^+)^T = \mathbf{B}\mathbf{B}^+; \quad \text{and} \quad (\mathbf{B}^+\mathbf{B})^T = \mathbf{B}^+\mathbf{B}.$$

A useful secondary property of \mathbf{B}^+ is that $\mathbf{z} = \mathbf{B}^+\mathbf{c}$ is the shortest length solution to the least squares problem $\mathbf{B}\mathbf{z} \approx \mathbf{c}$.

and

$$\mathbf{a}_{3 \times 1} = \begin{bmatrix} 0 \\ -g \\ 0 \end{bmatrix}. \quad (16)$$

Accordingly,

$$\mathbf{M}_{3 \times 3}^{-1/2} = \begin{bmatrix} m^{-1/2} & 0 & 0 \\ 0 & m^{-1/2} & 0 \\ 0 & 0 & m^{-1/2} \end{bmatrix}. \quad (17)$$

Second, we write out all the constraint equations. In this case, to keep the point on the sphere, the only constraint is

$$x_1^2 + x_2^2 + x_3^2 = l^2. \quad (18)$$

Third, we get the linear restriction on the acceleration vector by differentiating once or twice the constraints with respect to t . In this example, one differentiation of the constraint Eq. (18) gives

$$x_1 \dot{x}_1 + x_2 \dot{x}_2 + x_3 \dot{x}_3 = 0, \quad (19)$$

and one more differentiation results in

$$x_1 \ddot{x}_1 + x_2 \ddot{x}_2 + x_3 \ddot{x}_3 = -(\dot{x}_1^2 + \dot{x}_2^2 + \dot{x}_3^2). \quad (20)$$

Notice that Eq. (20), together with initial conditions on the displacements and velocities, describes the same constraint on the system as Eq. (18). In matrix form, Eq. (20) is

$$\begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} \ddot{x}_1 \\ \ddot{x}_2 \\ \ddot{x}_3 \end{bmatrix} = -(\dot{x}_1^2 + \dot{x}_2^2 + \dot{x}_3^2). \quad (21)$$

Let

$$\mathbf{A}_{1 \times 3} = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}, \quad (22)$$

$$\mathbf{\ddot{x}}_{3 \times 1} = \begin{bmatrix} \ddot{x}_1 \\ \ddot{x}_2 \\ \ddot{x}_3 \end{bmatrix}, \quad (23)$$

and

$$\mathbf{b}_{1 \times 1} = -(\dot{x}_1^2 + \dot{x}_2^2 + \dot{x}_3^2). \quad (24)$$

Eq. (21) then is reduced to

$$\mathbf{A} \mathbf{\ddot{x}} = \mathbf{b}, \quad (25)$$

which is the required form for the constraints in applying the basic formula (14). Since

$$\mathbf{A} \mathbf{M}_{1 \times 3}^{-1/2} = m^{-1/2} \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}, \quad (26)$$

it follows that

$$(\mathbf{A}\mathbf{M}^{-1/2})^+ = \frac{1}{m^{-1/2}(x_1^2 + x_2^2 + x_3^2)} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}. \quad (27)$$

Lastly, we substitute Eqs. (16), (17), (27), (24) and (22) into Eq. (14). This leads to

$$\begin{aligned} \ddot{\mathbf{x}} &= \mathbf{a} + \mathbf{M}^{-1/2}(\mathbf{A}\mathbf{M}^{-1/2})^+(\mathbf{b} - \mathbf{A}\mathbf{a}) \\ &= \begin{bmatrix} 0 \\ -g \\ 0 \end{bmatrix} + \frac{1}{x_1^2 + x_2^2 + x_3^2} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \left(-(\dot{x}_1^2 + \dot{x}_2^2 + \dot{x}_3^2) - \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} 0 \\ -g \\ 0 \end{bmatrix} \right) \\ &= \begin{bmatrix} 0 \\ -g \\ 0 \end{bmatrix} + \frac{g x_2 - (\dot{x}_1^2 + \dot{x}_2^2 + \dot{x}_3^2)}{x_1^2 + x_2^2 + x_3^2} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \end{aligned} \quad (28)$$

which can be verified to be the correct equation of motion for the spherical pendulum by using Lagrange's equations of motion of the first kind or otherwise.

4. CONCLUDING REMARKS. In this note, we have indicated the close connection between two cultures that are normally viewed as being quite separate: that of analytical mechanics and that of generalized inverses of matrices. Contributions should flow freely from one to the other. In particular, there are pedagogical implications for the teaching of generalized inverses early in the undergraduate curriculum. Many analytical and computational aspects of the generalized inverse form of the equations of constrained motion can be considered now [4]. Automation of the entire process, including machine evaluation of needed partial derivatives [5] and machine evaluation of the generalized inverses, should be both challenging and rewarding.

REFERENCES

1. Kalaba, R. and Udawadia, F. (1992) "On Constrained Motion," *Applied Mathematics and Computation*, Vol. 51, pp. 85-86.
2. Whittaker, E. (1944) *A Treatise on the Analytical Dynamics of Particles and Rigid Bodies: with an Introduction to the Problem of Three Bodies*, Dover Publications, New York, New York.
3. Lawson, C. and Hanson, R. (1974) *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, New Jersey.
4. Udawadia, F. and Kalaba, R. (1995) *A New Analytical Dynamics*, Forthcoming, Cambridge University Press, Cambridge.
5. Kagiwada, H., Kalaba, R., Rasakhoo, N., and Spingarn, K. (1986) *Numerical Derivatives and Nonlinear Analysis*, Plenum Press, New York, New York.

*Departments of Biomedical
and Electrical Engineering
University of Southern California
Los Angeles, CA 90089*

*School of Urban and Regional
Planning
University of Southern California
Los Angeles, CA 90089*

THE COMPUTER SCIENCE SAMPLER

Edited by: Catherine C. McGeoch

Off to the Races

Jeffrey Ondich

You are seated in a restaurant, about to eat some soup. You pick up your spoon. At that moment, crashing dishes interrupt your concentration. While you look away, your hungry eating companion slips the spoon from your hand. A moment later, you continue your task where you left off, and plunge your hand into the soup.

Most human beings, even mathematicians, would notice the absence of the spoon. But computer programs tend to be less adaptable than their authors. In this column, we will take a look at the trouble that can arise when programs that can be interrupted at any time try to share resources and communicate with one another.

Processes and Interrupts. On a computer system with one central processing unit (CPU), only one program can actually run at a time. A *pre-emptive multi-tasking* system creates the illusion of many programs running simultaneously by forcing the programs to take turns. One of the many jobs of a computer's operating system is to schedule *processes*. A process is a program in the midst of execution. At any given time, exactly one process is *running*—that is, its instructions are being executed by the CPU. Other processes are *ready* to run, and still others are *blocked* or *asleep*. A blocked process will not be given control of the CPU until some condition is met, such as the arrival of data the process requested from a disk.

Once a process is given control of the CPU, how does any other process get control back? One way is for the running process to put itself to sleep. For example, if a process requests some data from a hard disk, the process will block while waiting, and it will be up to the operating system to wake the process up when the information arrives. If a process doesn't give up the CPU voluntarily, it has to be forced out. Every so often, typically on the order of once per 60th of a second, a small quartz clock sends the CPU an *interrupt* (a term that beautifully exemplifies computer scientists' fondness for using a verb when a perfectly good noun is available). The interrupt causes the CPU to save the running process and give control to the operating system, which may then give control to any process that is ready to run.

Race conditions. Multi-tasking allows computer users to run many programs at once, and allows the CPU to spend most of its time doing useful work. But the pre-emption and inter-process communication that often come with multi-tasking

can cause trouble, too. Consider the bakery in the restaurant where you had your soup. Every day, the bakery sells 100 chocolate cakes by taking orders via the Internet. When a purchase order comes in, a new process is created to handle the order. All such processes have access to a variable named *cakesSold*, which is initialized to 0 every morning. This is what these processes do:

```
If cakesSold < 100,  
    Add 1 to cakesSold  
    Record the transaction  
    Send the order to the delivery truck.  
Otherwise,  
    Ask the customer to try again tomorrow.
```

Now suppose *cakesSold* = 99, and two orders arrive in quick succession, prompting the creation of processes A and B. Consider the following sequence of events.

1. A determines that *cakesSold* < 100
2. An interrupt occurs, and B gets control
3. B determines that *cakesSold* < 100
4. B adds 1 to *cakesSold*
5. B finishes its work, and A gets control
6. A adds 1 to *cakesSold*...

The value of *cakesSold* is now 101, so someone is going to have to bake another cake. Process A has, so to speak, failed to notice its spoon is missing. The problem here is that it takes more than one step to test and alter *cakesSold*, but interrupts can occur any time, leading to *race conditions*—trouble caused by two processes racing to use a shared resource.

Semaphores. To correct the problem with the cakes, we need some way to give Process A exclusive access to *cakesSold* for a short time. One way to do this involves *semaphores*, first suggested by Edsger Dijkstra in [1]. A semaphore has a non-negative integer value and a collection of processes, all blocked, that are said to be “sleeping on the semaphore.” Initially, a semaphore has no processes sleeping on it. Any process can perform either of two operations on a semaphore:

DOWN: If the semaphore’s value is greater than 0, the value is decreased by 1. If the value is 0, the value remains 0, and the process performing the DOWN goes to sleep and gets added to the semaphore’s collection of sleeping processes.

UP: If the semaphore’s value is 0 and its collection of sleeping processes is non-empty, one of those sleepers is awakened. Otherwise, the semaphore’s value is increased by 1.

DOWN and UP can be built into the operating system in such a way that they are *atomic*. That is, like a single machine language instruction, they cannot be interrupted.

Let’s see how a semaphore can be used to solve our cake problem. We will follow operating systems literature tradition and call our semaphore “mutex,” in reference to the mutually exclusive use of *cakesSold* we are trying to enforce. We will set the value of *mutex* to 1, which will mean that no process is using *cakesSold*

at the moment. Our order-handling procedure now looks like this:

```
DOWN(mutex)
If cakesSold < 100,
    Add 1 to cakesSold
    UP(mutex)
    Record the transaction
    Send the order to the delivery truck
Otherwise,
    UP(mutex)
    Ask the customer to try again tomorrow
```

Suppose A and B are using this new semaphore-based approach, and B interrupts A as before:

1. A does a DOWN on mutex, whose value becomes 0
1. A determines that *cakesSold* < 100
2. An interrupt occurs, and B gets control
3. B does a DOWN on mutex, and thus goes to sleep
4. A gets control, and adds 1 to *cakesSold*
5. A does an UP on mutex, waking B up

Even if B takes control again now, the danger is past. Process A had exclusive access to *cakesSold* long enough to test it and add 1.

Sending Messages. Semaphores are easy to misuse, and they do not provide a very abstract way of thinking about interprocess communication. A more comfortable abstraction is the notion of *message passing*. We would like process A to be able to send messages to and receive messages from process B. Using a message passing system, the cake problem could be solved by requiring processes to request permission from one another before altering *cakesSold*. Such an approach would be metaphorically more pleasing than the use of semaphores, and thus more reliable.

If our operating system provides us with semaphores, we can use them to implement a message passing system. In a sense, we can use the semaphores as a programming language, and write message-passing programs in that language.

Suppose for simplicity that only two processes, still called A and B, will be sending messages to one another. We will set aside memory, shared by A and B, to hold a “mailbox” for each process. Each mailbox will have enough space to hold exactly one message. We will also associate with A’s mailbox two semaphores, called *fullA* and *emptyA*, whose values will initially be 0 and 1, respectively. B’s mailbox will have two similarly named semaphores. From these pieces, we can construct the basic operations *SEND* and *RECEIVE*.

When B wants to SEND a message to A, here is what B will do:

```
DOWN(emptyA)
Write the message into A’s mailbox
UP(fullA)
```

To RECEIVE a message from B, A does the following:

```
DOWN(fullA)
Read the message in A’s mailbox
UP(emptyA)
```

Any time before B has finished SENDing, *fullA* will have value 0, so A will go to sleep if it tries to RECEIVE. Similarly, if B tries to SEND a second message before A has RECEIVE'd the first, B's DOWN on *emptyA* will put B to sleep. Thus, the problem of B and A trying to SEND and RECEIVE at the same time will not take place, nor will the problem of B SENDing too many messages for A's mailbox to hold.

Even though semaphores and message-passing systems can prevent race conditions, they can also be misused, causing a related problem called *deadlock*. In the restricted, two-process world of the examples above, both A and B might invoke SEND twice before either invoked RECEIVE, or both might perform a DOWN on a semaphore whose value was already zero. Either way, A and B would be asleep, each waiting for a wake-up call from the other.

Equivalence. Semaphores and message passing are not the only abstractions available for regulating interprocess communication. There are *event counters*, *monitors*, *rendezvous*, *sequencers*, *path expressions*, *serializers*, and more (see [2] for details). Each is just the right abstraction for some communication task. But, as it turns out, they are all equivalent.

We have already seen how semaphores can be used to implement a message passing system. To show that semaphores and message passing are equivalent, we need to construct semaphores from a message passing system.

Suppose we have atomic operations SEND and RECEIVE for message-passing, but our operating system provides us with no semaphores. To mimic semaphore behavior, we can create a special process—the semaphore Boss, if you will—to coordinate calls to UP and DOWN. For each semaphore, the Boss will maintain a value and a list of sleeping processes.

To perform an UP on a semaphore S, process A will SEND to the Boss a message containing the identity of the semaphore in question and a marker to indicate that A wants to do an UP. (To do a DOWN, A sends the Boss the same message with a different marker.) Then A invokes RECEIVE, and waits for a return message from the Boss. The content of the return message is irrelevant; what matters is that A will sleep until the message arrives.

The following procedures describe the actions of the Boss upon receiving a DOWN(S) message from A:

```
    If the value of S is greater than 0
        Subtract 1 from the value of S
        SEND a message to A
    Otherwise,
        Add A to the list of processes sleeping on S
```

and an UP(S) message from A:

```
    If the value of S is greater than 0
        Add 1 to the value of S
        SEND a message to A
    Otherwise,
        Remove one of the sleepers from the list
        of processes sleeping on S, and send that
        process a message
```

All semaphore operations are performed while the Boss is running, and so the unpredictable order of events that can lead to race conditions will not be a problem here.

References. Mutual exclusion, semaphores, message passing, and deadlock are discussed in virtually every elementary operating systems textbook. Two good examples are [5] and [4].

Dijkstra's original paper [1] is worth reading to understand the historical context in which semaphores were proposed. If your operating system doesn't even give you semaphores, you can still enforce mutual exclusion using the CPU-hogging technique of *busy waiting* (see [3]).

REFERENCES

1. E. W. Dijkstra, "Co-operating Sequential Processes," 1965, reprinted in *Programming Languages*, Genuys, F. (Ed.), Academic Press, 1968.
2. M. Maekawa, A. Oldehoeft, and R. Oldehoeft, *Operating Systems, Advanced Concepts*, Benjamin Cummings, 1987.
3. G. L. Peterson, "Myths about the Mutual Exclusion Problem," *Information Processing Letters*, v. 12, pp. 115-116, June 1981.
4. W. Stallings, *Operating Systems*, Prentice Hall, 1995.
5. A. S. Tanenbaum, *Modern Operating Systems*, Prentice Hall, 1992.

Department of Mathematics & Computer Science
Carleton College
Northfield, MN 55057
jondich@carleton.edu

Remote from human passions, remote even from the pitiful facts of nature, the generations have gradually created an ordered cosmos, where pure thought can dwell as in its natural home and where one, at least, of our nobler impulses can escape from the dreary exile of the actual world.

—Russell

THE EVOLUTION OF . . .

Edited by Abe Shenitzer

Mathematics, York University, North York, Ontario M3J1P3, Canada

Elliptic Curves

John Stillwell

In recent years, elliptic curves have played a leading role in number theory, most famously in Wiles' program to prove Fermat's last theorem. However, since these developments are highly technical, it may be useful to look back to earlier times, when elliptic curves led a simpler life. For about 1500 years, from the time of Diophantus to Newton, elliptic curves were known only as curves defined by certain cubic equations. This put them just a step beyond the conic sections, and some of their geometric and arithmetic properties can in fact be viewed as generalisations of properties of conics. In particular, it is possible to find rational solutions of both quadratic and cubic equations by simple geometric constructions.

It was only with the development of calculus, in the 17th century, that sharp differences between conics and elliptic curves began to emerge. Conic sections can be parametrised by rational functions. For example, the circle $x^2 + y^2 = 1$ is parametrised by

$$x = \frac{1 - t^2}{1 + t^2}, y = \frac{2t}{1 + t^2}$$

but the elliptic curves cannot. Their simplest parametrising functions are *elliptic functions*, which arise in calculus as the inverses of elliptic integrals, so-called because a typical example is the integral for the arc length of the ellipse. It is for this fairly accidental reason that they are called elliptic curves—an unfortunate accident since the ellipse itself is *not* an elliptic curve.

The difference between conics and elliptic curves was “felt” in the 17th century in the apparent intractability of elliptic integrals, though the parametrisation of cubic curves was not known at that time. The idea of inverting elliptic integrals to create elliptic functions had to wait until the early 19th century. The nonrationality of elliptic curves was not fully understood until the mid-19th century, when the introduction of complex coordinates revealed a *topological* difference between them and conics. This brings us within sight of the modern view of elliptic curves—a remarkable synthesis of number theory, geometry, algebra, analysis and topology. In what follows I shall attempt to describe what led up to this state of affairs.

Diophantus. Very little is known about Diophantus except that he lived sometime between 150 AD and 350 AD and was a wizard at finding rational solutions to polynomial equations in two or more variables. His *Arithmetica* (available in the

English edition of Heath [4]), contains the solutions of hundreds of equations, among them the following instructive examples.

1. A rational solution of $x^2 + y^2 = 16$, other than an obvious one such as $x = 0$, $y = 4$, is found by solving the simultaneous equations

$$\begin{aligned}x^2 + y^2 &= 16, \\ y &= 2x - 4,\end{aligned}$$

which yield the solution $x = 16/5$, $y = 12/5$ (Heath [4], p. 145).

2. A rational solution of $x^3 - 3x^2 + 3x + 1 = y^2$, other than the obvious one $x = 0$, $y = 1$, is found by solving the simultaneous equations

$$\begin{aligned}x^3 - 3x^2 + 3x + 1 &= y^2, \\ y &= \frac{3}{2}x + 1,\end{aligned}$$

which yield the solution $x = 21/4$, $y = 71/8$ (Heath [3], p. 242).

How did Diophantus choose the linear equations in these two examples? The simplest explanation is geometric, although he makes no mention of geometry.

In the first example the linear equation represents a line through the “obvious” rational point $(0, 4)$. Its slope is not important, since any line through $(0, 4)$ with rational slope t will meet the circle at a second rational point $(8t/(1 - t^2), (4t^2 - 4)/(1 + t^2))$. Conversely, all rational points on the circle are obtainable in this way, so Diophantus has essentially *parametrised* the rational points on the circle by rational functions of a rational parameter t .

The linear equation in the second example has an even stronger geometric smell. It is the *tangent* to $x^3 - 3x^2 + 3x + 1 = y^2$ at the “obvious” rational point $(0, 1)$. Here there is no option about the slope because a line has to meet a cubic curve in *two* rational points for its third intersection to be rational. When only one rational point is known, this forces us to use the tangent, which is the line through two “coincident” points.

It is possible, of course, that Diophantus discovered these facts purely algebraically, and did not notice their geometric interpretation. However, that would be a truly amazing departure from the Greek mathematical culture of his time. Even in the more algebraic culture of the 17th century, Fermat and Newton immediately recognised Diophantus’ work as geometry, with Newton [6] explicitly interpreting Diophantus’ solutions as chord and tangent constructions. Later discoveries added more weight to the geometric interpretation, as we shall see below.

Fermat and Newton. Fermat was the first mathematician to make significant progress in number theory beyond Diophantus. Among his many discoveries were methods for proving *nonexistence* of integer or rational solutions for certain equations. For example, he proved that there are no positive rationals a, b, c such that

$$a^4 \pm b^4 = c^2$$

This implies in particular that no positive integer fourth powers sum to a fourth power (the $n = 4$ case of Fermat’s last theorem), but it is also a statement about an elliptic curve. It says that there are no nontrivial rational points on the curve

$$y^2 = 1 - x^4,$$

since a rational point $(p/r, q/r)$ with $p, q \neq 0$ and

$$\frac{p^2}{r^2} = 1 - \frac{q^4}{r^4}$$

gives nonzero integers $a = r, b = q, c = pr$ with $a^4 - b^4 = c^2$.

Now I know I said that elliptic curves are cubics, but they are cubic *in a suitable coordinate system*. Any quartic curve of the form

$$y^2 = (x - \alpha)(x - \xi)(x - \gamma)(x - \delta)$$

can be rewritten

$$\left(\frac{y}{x - \alpha}\right)^2 = \left(1 - \frac{\beta - \alpha}{x - \alpha}\right)\left(1 - \frac{\gamma - \alpha}{x - \alpha}\right)\left(1 - \frac{\delta - \alpha}{x - \alpha}\right)$$

and hence it is cubic in the coordinates

$$X = \frac{1}{x - \alpha}, Y = \frac{y}{x - \alpha^2}.$$

In particular, $y^2 = 1 - x^4$ is a cubic $Y^2 = 4X^3 - 6X^2 + 4X - 1$ in the coordinates $X = 1/(1 - x), Y = y/(1 - x)^2$. Notice that this is an appropriate coordinate change from the point of view of number theory, because it makes the rational points (x, y) on one curve correspond to the rational points (X, Y) on the other. Such a coordinate change is called *birational*.

Newton made the surprising discovery that all cubic equations in x and y can be reduced to the form

$$Y^2 = X^3 + aX + b$$

by a birational coordinate transformation. In fact, the transformations he used were simply projections. He called this “genesis of curves by shadows”. His result can be viewed as an analogue of the well known theorem that second degree curves are conic sections and hence, in nondegenerate cases, projections of the circle. The degenerate cubic curves are those for which the right hand side $X^3 + aX + b$ has a repeated factor. The corresponding repeated root $X = \alpha$ is either a double point (Fig. 1) or cusp (Fig. 2) of the curve, and by drawing a line of slope t through this point we obtain the coordinates of the general point on the curve as rational functions of t .

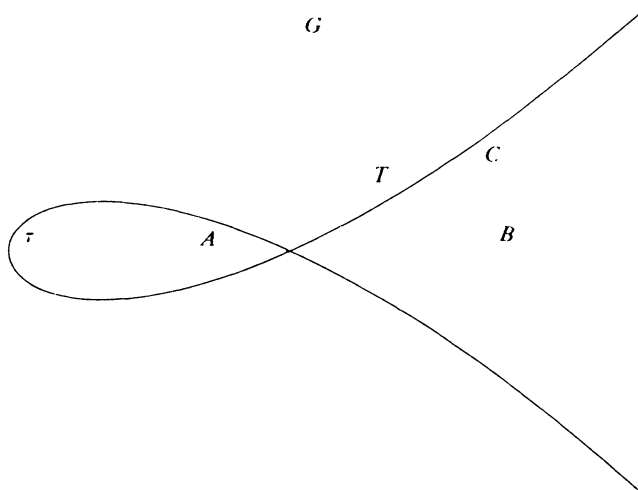


Figure 1. Cubic with double point.

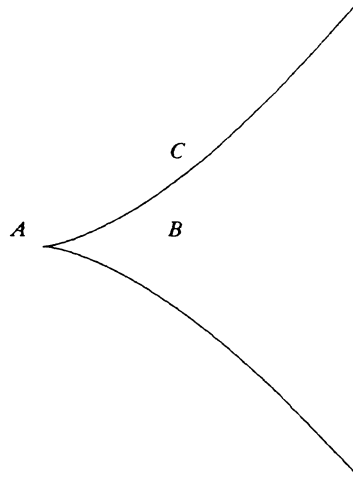


Figure 2. Cubic with cusp.

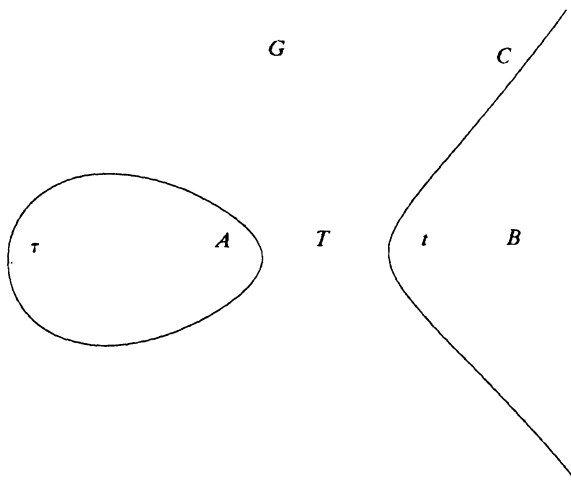


Figure 3. Nonsingular cubic.

The curves for which $X^3 + aX + b$ has no repeated factor cannot be parametrised by rational functions, and are what we now call elliptic curves (Fig. 3).

Elliptic integrals. Early in the development of integral calculus, mathematicians encountered the problem of “rationalising” square roots of polynomials. For example, to find the area or arc length of a circle one finds an integral involving $\sqrt{1-x^2}$. This can be rationalised by the “Diophantine” substitution $x = (1-t^2)/(1+t^2)$, and fact Jakob Bernoulli [1], in a similar situation, actually attributed the substitution to Diophantus. He used it to obtain the expression

$$\frac{\pi}{4} = \int_0^1 \frac{dt}{1+t^2},$$

whence he obtained the famous series

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots$$

by expanding $1/(1+t^2)$ in a geometric series and integrating term by term.

Integrals involving square roots of cubic or quartic polynomials proved more intractable. They were called *elliptic integrals* because one of them expresses the arc length of the ellipse. Cubics and quartics were lumped together because of birational equivalences between them, as noted above for $y^2 = 1 - x^4$ and $Y^2 = 4X^3 - 6X^2 + 4X - 1$. Such integrals arise from a great number of natural geometric and mechanical problems, so a lot of effort was expended on them, but without success.

Perhaps the first to see why rationalisation might be impossible was Jakob Bernoulli [2], who noted that a rationalisation of $\sqrt{1-x^4}$, at least by a rational function $x = f(t)$ with rational coefficients, would violate Fermat's theorem on the nonexistence of positive integer solutions of $a^4 \pm b^4 = c^2$. In fact, it can be shown that $\sqrt{1-x^4}$ cannot be rationalised by any rational function $x = f(t)$, by repeating Fermat's argument with polynomials in place of integers, so Jakob Bernoulli was on the right track. However, this type of argument was not used until the 19th century, so the nature of elliptic integrals remained unclear until then (when ideas not only from number theory, but also from analysis and topology, were directed at the problem).

Elliptic functions. In the 1820s, Abel and Jacobi finally saw what to do with elliptic integrals—*Invert* them. Instead of studying the integral

$$u = g^{-1}(x) = \int_0^x \frac{dt}{\sqrt{t^3 + at + b}},$$

say, study its inverse function $x = g(u)$. The gain in simplicity is comparable to studying the function $x = \sin u$ instead of the integral $\sin^{-1} x = \int_0^x (dt/\sqrt{1-t^2})$. In particular, instead of a multi-valued integral $g^{-1}(x)$, one has a *periodic function* $x = g(u)$.

The difference between $\sin u$ and $g(u)$ is that the periodicity of $g(u)$ cannot be properly seen until complex values of the variables are admitted, at which stage it emerges that $g(u)$ has *two* periods. That is, there are nonzero $\omega_1, \omega_2 \in \mathbb{C}$, with $\omega_1/\omega_2 \notin \mathbb{R}$, such that

$$g(u) = g(u + \omega_1) = g(u + \omega_2).$$

The two periods can be brought to light in various ways. One method, originating with Eisentein [1847] and commonly used today, is to write down a function that obviously has periods ω_1 and ω_2 , namely

$$g(u) = \sum_{m,n \in \mathbb{Z}} \frac{1}{(u + m\omega_1 + n\omega_2)^2},$$

and derive its properties by manipulation of infinite series. Eventually one finds that $g^{-1}(x)$ is an integral of the type we started with.

A more insightful approach though harder to make rigorous, is to study the behaviour of the integrand $1/\sqrt{t^3 + at + b}$ as t varies over the complex plane. Following Riemann [7], and viewing the 2-valued “function” $1/\sqrt{t^3 + at + b}$ as a 2-sheeted surface over \mathbb{C} , one finds that there are two independent closed paths of

integration, over which the integrals are ω_1 and ω_2 . This accounts for the periods ω_1 and ω_2 of the inverse function $g(u)$.

Since $g(u) = x$, it follows by basic calculus that

$$g'(u) = \frac{dx}{du} = \frac{1}{du/dx} = \frac{1}{1/\sqrt{x^3 + ax + b}} = \sqrt{x^3 + ax + b} = y,$$

so $x = g(u)$, $y = g'(u)$ gives a parametrisation of the curve $y^2 = x^3 + ax + b$. With a little more work it can be shown that $u \mapsto (g(u), g'(u))$ is in fact a continuous one-to-one correspondence between $\mathbb{C}/\langle \omega_1, \omega_2 \rangle$ and the curve. $\mathbb{C}/\langle \omega_1, \omega_2 \rangle$ is the quotient of \mathbb{C} by the subgroup generated by ω_1 and ω_2 and is topologically a *torus*, hence so is the curve $y^2 = x^3 + ax + b$. This is the deeper reason why elliptic curves are not rationally parametrisable—a curve parametrised by rational functions $x = p(u)$, $y = q(u)$ is the topological image of the completed plane $\mathbb{C} \cup \{\infty\}$ of u values, and $\mathbb{C} \cup \{\infty\}$ is topologically a *sphere*.

Another consequence of the parametrisation $x = g(u)$, $y = g'(u)$ is that the curve $y^2 = x^3 + ax + b$ is an abelian group. The “sum” of points with parameter values u_1, u_2 is simply the point with parameter value $u_1 + u_2$. Under this definition of sum, the curve is isomorphic to the group $\mathbb{C}/\langle \omega_1, \omega_2 \rangle$. Amazingly, there is an equivalent definition of the sum that Diophantus would have understood (and which helps to explain why elliptic functions are useful in number theory): the sum of the points P_1 and P_2 is simply the reflection, in the x -axis, of the third point on the curve collinear with P_1 and P_2 (Fig. 4). For an explanation

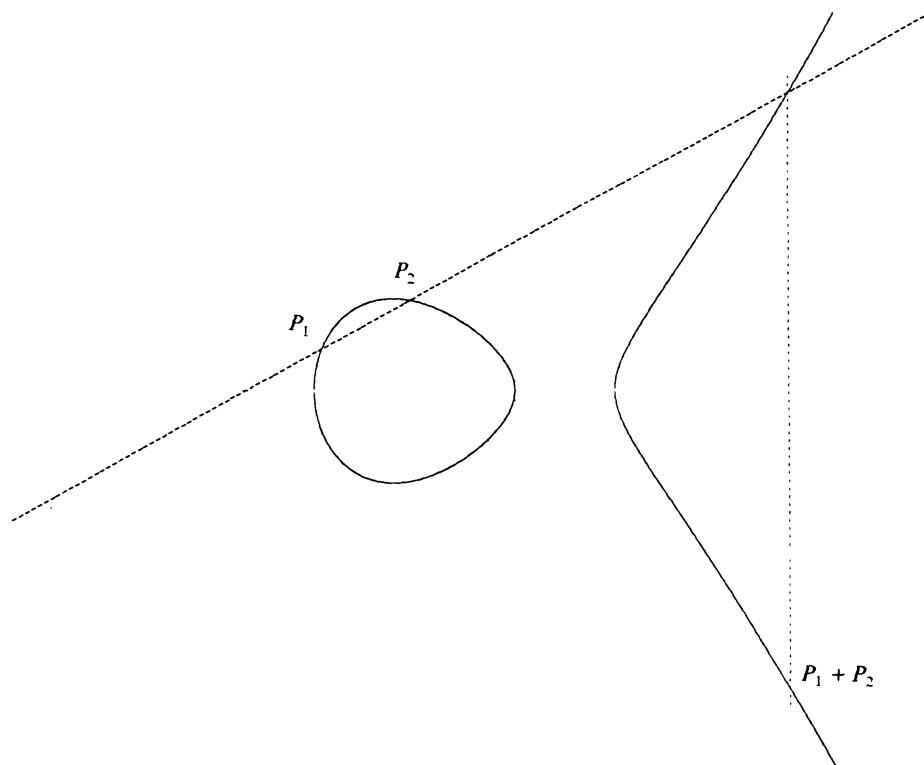


Figure 4. Addition of points on an elliptic curve (from Koblitz [5]).

of this face we must refer the reader to a recent book on elliptic curves, such as Koblitz [5]. In the same book you will find many beautiful modern results on elliptic curves, motivated by ancient problems in number theory and geometry.

REFERENCES

1. Bernoulli, Jakob (1696) Positionum de seriebus infinitis pars tertia. *Werke*, 4, 85–106.
2. Bernoulli, Jakob (1704) Positionum de seriebus infinitis ... pars quinta. *Werke*, 4, 127–147.
3. Eisenstein, G. (1847) Beiträge zur Theorie der elliptischen Functionen. *J. reine angew. Math.* 35, 137–274.
4. Heath, T. L. (1910) *Diophantus of Alexandria*, Cambridge University Press.
5. Koblitz, N. (1985) *Introduction to Elliptic Curves and Modular Forms*, Springer-Verlag, New York.
6. Newton, I. (late 1670s) De resolutione quaestionum circa numeros. *Math. Papers* 4, 110–115.
7. Riemann, G. B. H. (1851) Grundlagen für eine allgemeine Theorie der Functionen einer veränderlichen complexen Grösse. *Werke*, 2nd ed., 3–48.

Department of Mathematics
Monash University
Clayton 3168
AUSTRALIA
stillwell@monash.edu.au.

Without the concepts, methods and results found and developed by previous generations right down to Greek antiquity one cannot understand either the aims or the achievements of mathematics in the last fifty years.

—H. Weyl (in 1950)

PROBLEMS AND SOLUTIONS

Edited by:

Richard T. Bumby, Fred Kochman and Douglas B. West

Proposed problems should be sent to the MONTHLY PROBLEMS address given on the inside front cover. Please include solutions and relevant references. Three copies of all items needed to evaluate the problem should be sent.

Solutions of published problems should arrive at the MONTHLY PROBLEMS address given on the inside front cover before April 30, 1996. If possible, solutions should be typed with double spacing. Two copies suffice. Several solutions may be mailed together, but they should be on separate sheets of paper. The problem number and the solver's name and mailing address should appear on each solution. A mailing label should be included if an acknowledgment is desired.

The published solution is likely to be based on a solution that is complete and correct. Additional information, such as references to other appearances of the problem or its solution, is also welcome.

An asterisk () after the number of a problem, or part of a problem, indicates that no solution is currently available.*

PROBLEMS

10480. *Proposed by Shmuel Rosset, Tel Aviv University, Ramat Aviv, Israel.*

The standard proof of Cayley's theorem shows that S_n , the symmetric group of degree n , contains as subgroups every group of order n . Which groups of order n are contained in A_n , the alternating group of degree n ?

10481. *Proposed by Frank Schmidt, Arlington, VA.*

Let $f(n)$ denote the number of n by n matrices whose entries are 0 or 1 that are positive semi-definite.

Let $g(n)$ denote the number of n by n matrices whose entries are 0 or 1 that are positive definite.

Evaluate $f(n)$ and $g(n)$.

10482. *Proposed by Emre Alkan (student), Bosphorus University, İstanbul, Turkey, and Murray S. Klamkin, University of Alberta, Edmonton, Alberta, Canada.*

Given a regular n -gonal pyramid with apex P and base $A_1 A_2 \dots A_n$, denote $\angle A_i P A_{i+1}$ by α with $0 < \alpha \leq 2\pi/n$. If points B_i are chosen on the rays PA_i ($i = 1, 2, \dots, n$), determine the maximum and minimum values of

$$\frac{|PB_1| + |PB_2| + \dots + |PB_n|}{|B_1 B_2| + |B_2 B_3| + \dots + |B_n B_1|}.$$

10483. *Proposed by Stanley Rabinowitz, Westford, MA.*

Given an odd positive integer n , let A_1, A_2, \dots, A_n be a regular n -gon with circumcircle Γ . A circle O_i with radius r is drawn externally tangent to Γ at A_i for $i = 1, 2, \dots, n$. Let P be any point on Γ between A_n and A_1 . A circle C (with any radius) is drawn externally tangent to Γ at P . Let t_i be the length of the common external tangent between the circles C and O_i . Prove that

$$\sum_{i=1}^n (-1)^i t_i = 0.$$

10484. *Proposed by N. Bebiano and J. da Providência, Universidade de Coimbra, Coimbra, Portugal.*

Let $n \geq 3$, and let $\alpha = (\alpha_1, \dots, \alpha_n)$ and $\beta = (\beta_1, \dots, \beta_n)$ be complex row vectors such that $\{\alpha_1, \dots, \alpha_n, \beta_1, \dots, \beta_n\}$ are all distinct. Consider the $n!$ complex numbers (counting multiplicities)

$$z_\sigma = \prod_{i=1}^n (\alpha_i - \beta_{\sigma(i)})$$

where σ runs through all permutations of $\{1, \dots, n\}$. Let $P(\alpha, \beta)$ denote the convex hull of the z_σ . Prove that $P(\alpha, \beta)$ is a line segment if and only if all the α_i and β_j lie on a common circle or straight line.

10485. *Proposed by David Bradley, Simon Fraser University, Burnaby, B. C., Canada.*

Find the real numbers r that satisfy the equation

$$\int_0^\infty \frac{dx}{(1+x^r)^r} = 1$$

for $r \in \mathbb{R}$ with $r > 1$.

10486. *Proposed by Joseph H. Silverman, Brown University, Providence, RI.*

Let $a, b > 0$ and $\alpha > 1$ be real numbers, and define a function

$$Z(s) = \sum_{n \in \mathbb{Z}} \frac{1}{(a\alpha^n + b\alpha^{-n})^s}$$

for $s \in \mathbb{C}$, $\Re(s) > 0$.

(a) Prove that $Z(s)$ has a meromorphic continuation to all of \mathbb{C} .

(b) Find the poles of $Z(s)$.

(c) Find the residues of $Z(s)$ at its poles.

SOLUTIONS

Positive Deformations of the Cauchy Matrix

10265[1992, 957]. *Proposed by Bjorn Poonen (student), University of California, Berkeley, CA.*

Let $a_1, \dots, a_n, b_1, \dots, b_n, \alpha$ be real numbers with b_1, \dots, b_n and α all positive. Prove

$$\sum_{i=1}^n \sum_{j=1}^n \frac{a_i a_j}{(b_i + b_j)^\alpha} \geq 0.$$

Solution I by Donald A. Darling, Newport Beach, CA. Let μ be a positive measure on $[0, \infty)$, and let its Laplace transform $f(s)$ have an abscissa of convergence s_0 : $f(s) = \int e^{-st} \mu(dt)$, $s > s_0$. If $b_i > s_0$, $i = 1, 2, \dots, n$ and a_i is real, $i = 1, 2, \dots, n$, then

$$\begin{aligned} \sum_{i,j=1}^n a_i a_j f(b_i + b_j) &= \sum_{i,j=1}^n a_i a_j \int e^{-(b_i+b_j)t} \mu(dt) \\ &= \int \left(\sum_{i=1}^n a_i e^{-b_i t} \right)^2 \mu(dt) \\ &\geq 0. \end{aligned}$$

In this problem, $\mu(dt) = \frac{1}{\Gamma(\alpha)} t^{\alpha-1} dt$ and $s_0 = 0$.

Solution II by David G. Cantor, Del Mar, CA. Note first that it suffices to consider the case in which the b_i are distinct, say $b_1 < b_2 < \dots < b_n$, since for every fixed choice of a_1, a_2, \dots, a_n , the given expression is continuous in the b_j .

We first prove

Lemma. Suppose that u_1, u_2, \dots, u_n are real numbers, not all zero. Then, the function

$$f(x) = \sum_{j=1}^n \frac{u_j}{(b_j + x)^\alpha}$$

has at most $n - 1$ positive zeros x .

Proof. Use induction on n . The result is clear when $n = 1$. Suppose that $n > 1$ and put

$$g(x) = (b_n + x)^\alpha f(x) = c_n + \sum_{j=1}^{n-1} u_j \left(\frac{b_n + x}{b_j + x} \right)^\alpha.$$

Then

$$\begin{aligned} g'(x) &= -\alpha \sum_{j=1}^{n-1} u_j \left(\frac{b_n + x}{b_j + x} \right)^{\alpha-1} \frac{b_n - b_j}{(b_j + x)^2} \\ &= -\alpha (b_n + x)^{\alpha-1} \sum_{j=1}^{n-1} u_j \frac{b_n - b_j}{(b_j + x)^{\alpha+1}} \end{aligned}$$

The inductive hypothesis tells us that $g'(x)$ has at most $n - 2$ positive zeros, so $g(x)$ and $f(x)$ have at most $n - 1$ positive zeros. This completes the proof of the lemma.

We turn now to the main result. First note that if A is the n by n matrix whose (i, j) entry is $1 / (b_i + b_j)^\alpha$, then $\det A \neq 0$. Indeed, if a nonzero linear combination of its columns were zero, say

$$\sum_{j=1}^n \frac{u_j}{(b_i + b_j)^\alpha} = 0$$

for $1 \leq i \leq n$, then the function $f(x)$ of the lemma would have n zeros b_1, b_2, \dots, b_n . This determinant is a nonzero continuous function of its parameters as long as they satisfy the specified inequalities. Thus, the sign for all $\alpha > 0$ can be found by looking at $\alpha = 1$. In this case, A is a Cauchy matrix whose determinant is

$$\frac{\prod_{i=1}^n \prod_{j=i+1}^n (b_i - b_j)(b_i - b_j)}{\prod_{i=1}^n \prod_{j=1}^n (b_i + b_j)(b_i + b_j)}$$

which is positive.

This shows that the determinant of the given quadratic form in the a_i and all of its principal subdeterminants are positive. A standard criterion for a symmetric matrix to be positive definite (see F. R. Gantmacher, *The Theory of Matrices*, Chelsea, 1960, vol. I, Theorem X.3, p. 306 or Roger A. Horn & Charles R. Johnson, *Matrix Analysis*, Cambridge, 1985, Theorem 7.2.5, p. 404) shows that

$$\sum_{i=1}^n \sum_{j=1}^n \frac{a_i a_j}{(b_i + b_j)^\alpha} > 0.$$

Editorial comment. Frank Schmidt noted that the solution could be extracted from Lemmas 5 & 6 of R. Bapat, "Multinomial probabilities, permanents and a conjecture of Karlin and Rinott", *Proc. Amer. Math. Soc.* 102 (1988), 467–472. These results may be summarized by the following statement.

Theorem. If $B = (b_{ij})$ is symmetric with real positive entries and precisely one positive eigenvalue, then $(1 / b_{ij}^\alpha)$ is positive semidefinite for all $\alpha > 0$

To apply the result in this case, take $B = (b_i + b_j)$. The matrix B has rank at most 2, allowing easy analysis of its spectrum.

The GCHQ Problem Solving Group gave an explicit computation, valid for general values of $\alpha > 0$, that is similar to the inductive proof of the characterization of positive definite matrices quoted in Solution II.

All other correct solutions were similar to Solution I, though usually restricted to the special values of μ and s_0 given at the end of that solution.

Solved also by R. J. Chapman (U. K.), N. D. Elkies, P. J. Fitzsimmons, G. Letac (France), O. P. Lossers (The Netherlands), A. D. Melas (Greece), R. Mercer (Canada), F. Schmidt, A. Tissier (France), L. Wertheim (student, Russia), GCHQ Problem Solving Group (U. K.), Western Maryland College Problems group, and the proposer. Four incorrect solutions were received.

The Tarry-Escott Problem

10284[1993, 185]. Proposed by Liang-shin Hahn, University of New Mexico, Albuquerque, NM.

For each positive integer l , show that there exists a positive integer n and a partition of $\{1, \dots, n\}$ as a disjoint union of two sets A and B , such that for $1 \leq i \leq l$,

$$\sum_{a \in A} a^i = \sum_{b \in B} b^i.$$

Solution by Hillel Gauchman and Ira Rosenholtz, Eastern Illinois University, Charleston, IL. We prove a more precise statement: For each non-negative integer l , there is a partition of $\{0, 1, \dots, 2^{l+1} - 1\}$ into sets A_l, B_l such that

$$\sum_{a \in A_l} a^j = \sum_{b \in B_l} b^j \quad \text{for } 0 \leq j \leq l.$$

We use the convention $0^0 = 1$ so that $\sum_{c \in C} c^0 = |C|$ even if $0 \in C$.

Let $A_0 = \{1\}$ and $B_0 = \{0\}$, and define $A_{n+1} = A_n \cup (2^{n+1} + B_n)$ and $B_{n+1} = B_n \cup (2^{n+1} + A_n)$, where $x + C = \{x + c : c \in C\}$. Trivially, A_0, B_0 have the desired property for $l = 0$. Proceeding by induction, we suppose that A_n, B_n have the desired property for $l = n$. Given an integer j with $0 \leq j \leq n + 1$, we compute

$$\begin{aligned} \sum_{a \in A_{n+1}} a^j - \sum_{b \in B_{n+1}} b^j &= \\ &= \left(\sum_{b \in B_n} (2^{n+1} + b)^j - \sum_{a \in A_n} (2^{n+1} + a)^j \right) - \left(\sum_{b \in B_n} b^j - \sum_{a \in A_n} a^j \right) \\ &= \sum_{k=0}^j \binom{j}{k} 2^{(n+1)(j-k)} \left(\sum_{b \in B_n} b^k - \sum_{a \in A_n} a^k \right) - \left(\sum_{b \in B_n} b^j - \sum_{a \in A_n} a^j \right) \end{aligned}$$

By the induction hypothesis, all terms are 0 when $0 \leq j \leq n$. When $j = n + 1$, the formula reduces to $(\sum_{b \in B_n} b^{n+1} - \sum_{a \in A_n} a^{n+1}) - (\sum_{b \in B_n} b^{n+1} - \sum_{a \in A_n} a^{n+1})$, which also equals 0.

Editorial comment. As Richard Guy commented, this is the Tarry-Escott problem, mentioned in L. E. Dickson, *History of the Theory of Numbers*, Chelsea, 1971, volume 2, 709-710. Also, A. E. Caicedo Núñez & J. C. Vera Lizcano located it as MONTHLY Problem E1312 [1958, 284; 1958, 776] by C. F. Pinzka. They also found it in Loren Larson, *Problem Solving Through Problems*, Springer, 1983, 163-164. Raul A. Simon found it in Joe Roberts, *Elementary Number Theory*, MIT, 1977, p. 88, 110S-111S.

The Tarry-Escott problem has a huge literature. All solvers showed that 2^{n+1} numbers suffice to provide simultaneous equalities up to the n th power. G. Myerson noted that an entire book has been devoted to the subject (A. Gloden, *Mehrgradige Gleichungen*, Noordhoff, 1944). Several solvers noted that the result extends to three or more sets of integers. This is Prouhet's problem, which preceded the Tarry-Escott problem, as E.M. Wright observed in "Prouhet's 1851 solution of the Tarry-Escott problem of 1910," this MONTHLY 66(1959), 199-201.

Solved also by 38 readers and the proposer.

A Recurrence Related to Counting Involutions

10347[1993, 951]. Proposed by T. S. Nanjundiah, University of Mysore, Mysore, India.

For integer $n \geq 1$, define real numbers R_n by

$$R_1 = 1 \quad R_{k+1} = 1 + \frac{k}{R_k} \quad (k \geq 1).$$

Prove that

$$\sqrt{n - \frac{3}{4}} + \frac{1}{2} \leq R_n \leq \sqrt{n + \frac{1}{4}} + \frac{1}{2}$$

for $n \geq 1$.

Editorial comment. All solutions followed the outline below.

Let $\alpha_n = (1 + \sqrt{4n+1}) / 2$. Since $f_n(x) = 1 + n/x$ is a decreasing function, the result will follow by induction from $\alpha_n = f_n(\alpha_n)$ and $\alpha_{n+1} \geq f_n(\alpha_{n-1})$. These results are easily verified.

In fact, as noted by Frank Schmidt and the National Security Agency Problems Group (independently), this result can be found in Leo Moser & Max Wyman, "On solutions of $x^d = 1$ in symmetric groups", *Canadian J. Math.* 7 (1955), 159–168.

John Gaisser and Jonathan Sorenson (jointly) suggested a generalization in which R_1 is arbitrary and subsequent R_k are defined by

$$R_{k+1} = R_1 + \frac{k}{R_k}.$$

Then, for $R_1 \geq 1$, one has

$$\sqrt{n-1 + \frac{R_1^2}{4}} + \frac{R_1}{2} \leq R_n \leq \sqrt{n + \frac{R_1^2}{4}}$$

for $n \geq 1$. They conjecture that a similar conclusion holds for $R_1 < 1$, provided that one takes $n \geq n_0(R_1)$, and they give the following sample values of n_0 : $n_0(9/10) = 4$, $n_0(1/2) = 32$, $n_0(1/5) = 410$.

H.-J. Seiffert wrote R_k in the form p_k/p_{k-1} with $p_0 = p_1 = 1$ and the p_n satisfying the recurrence $p_{k+1} = p_k + kp_{k-1}$. Again, an inductive argument is used. The recurrence for the p_k allows one to obtain a generating function

$$F(z) = \sum_{k=0}^{\infty} \frac{p_k}{k!} z^k = e^{z+(z^2/2)}.$$

This approach is also considered in the paper of Moser and Wyman cited above.

Solved by 56 readers and the proposer.

Collaborating editors: David F. Appleyard, Paul T. Bateman, Duane M. Broline, Barry W. Brunson, Frank S. Cater, Gulbank D. Chakerian, Underwood Dudley, Gerald A. Edgar, Michael A. Filaseta, Ira M. Gessel, Richard A. Gibbs, Jerrold R. Griggs, Douglas A. Hensley, John R. Isbell, Mourad E. H. Ismail, Murray Klamkin, Daniel J. Kleitman, Frederick W. Luttmann, Frank B. Miles, Richard Pfiefer, Stephen L. Portnoy, J. O. Shallit, John Henry Steelman, Kenneth B. Stolarsky, David E. Tepper, Douglas B. Tyler, Daniel Ullman, and William E. Watkins.

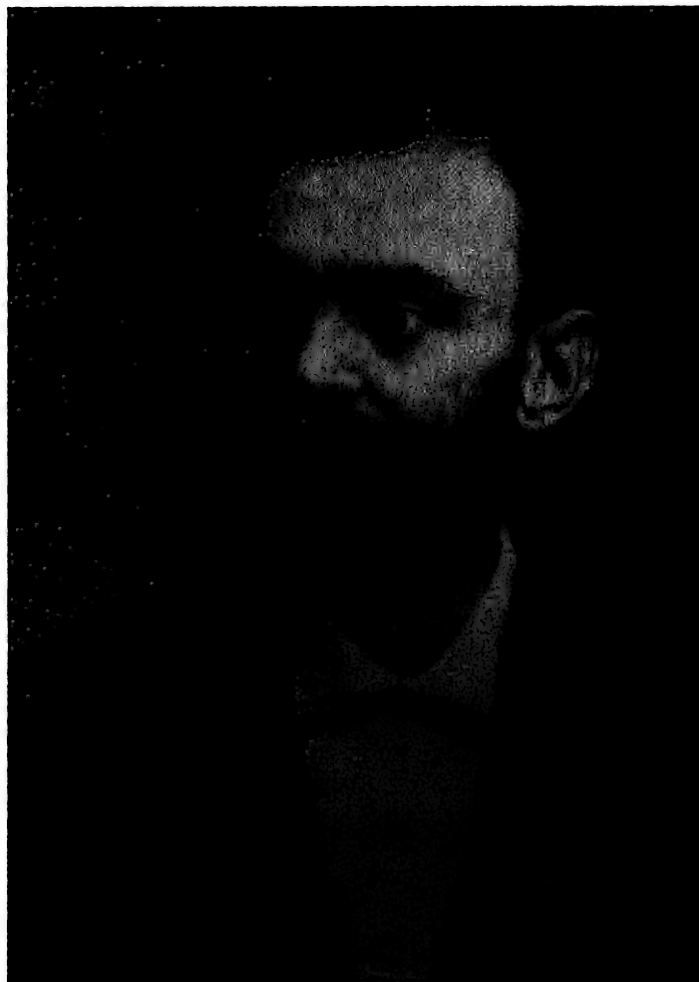
I have often observed ... that among some of the most capable, research-wise of new Ph.D.s can often be found the greatest lack of knowledge concerning the background and significance of their work, as well as abysmal ignorance of the reasons for doing it and of the general nature of mathematics. In fact, they are uneducated specialists.

—R. L. Wilder (1972)

The American Mathematical Monthly



Volume 102, Number 10 / DECEMBER 1995



Alfred Nobel
(see page 888)

NOTICE TO AUTHORS

The *Monthly* publishes articles, notes, and other features about mathematics and the profession. The readership of the *Monthly* is intended to include everybody who is mathematically inclined, including of course professional mathematicians and students of mathematics at all collegiate levels. While no single article or feature is likely to appeal to everyone, material should interest and be accessible to a large number of readers. This is the most important criterion for acceptance.

Articles may be expositions of old results or presentations of new ones. They may concern all of mathematics or one small area, a broad development or a single application, historical reminiscences or one important event. While some articles may contain the author's new research, the novelty of material and generally of the results is far less important than the clarity of exposition and general interest. Discussing one illuminating case of a well known result is far better than providing all the details of an obscure but new proposition. Articles in the *Monthly* are supposed to inform and to entertain; they are meant to be read rather than archived.

Notes are short and possibly informal articles. A note may concern a clever new proof of an old theorem, a novel way to present tired material, or a lively discussion of a philosophical (but still mathematical) issue. Also, any topic is suitable, so long as it is related to mathematics. Because a note is short, the first few sentences are the most important part: They should explain the purpose and invite the reader in. Photographs or diagrams often will attract the reader's attention.

All articles and notes should be sent to:

ROGER HORN
1515 Mineral Square, Room 142
University of Utah
Salt Lake City, UT 84112

Please send 3 copies, typewritten on only one side of the paper. Illustrations should be carefully drawn on separate sheets of paper in black ink; the original should be without lettering and two copies should have appropriate captions and lettering indicated.

Proposed problems or solutions should be sent to:

Richard BUMBY,
P.O. Box 10971
New Brunswick, NJ 08906-0971.

Please send 2 copies of all material, typewritten if possible.

Letters to the Editor, both for publication and for private reading, should be sent to the Editor at the address given above. Comments, including criticisms, are welcome, as are all suggestions for making the *Monthly* a lively, entertaining, and informative journal.

EDITOR:

JOHN H. EWING

ASSOCIATE EDITORS:

RONALD BOOK	JOAN HUTCHINSON
PETER BORWEIN	FRED KOCHMAN
RICHARD BUMBY	CATHERINE MCGEOCH
DENNIS DETURCK	RICHARD NOWAKOWSKI
UNDERWOOD DUDLEY	ARNOLD OSTEBEE
JOHN DUNCAN	LEERUBEL
JOAN FERRINI-MUNDY	LYNN STEEN
JOSEPH GALLIAN	STAN WAGON
STEVEN GALOVICH	DOUGLAS WEST
RICHARD GUY	HERBERT WILF
DARRELL HAILE	SANDY ZABELL
PAUL HALMOS	PAUL ZORN

EDITORIAL ASSISTANT:

MISTY CUMMINGS

STAFF ARTIST:

MIKE CAGLE

Reprint permission:

MARCIA P. SWARD, Executive Director

Advertising Correspondence:

Ms. ELAINE PEDREIRA, Advertising Manager

Subscription correspondence, change of address, and other inquiries:

Membership / Subscriptions Department

All at the address:

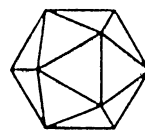
The Mathematical Association of America
1529 Eighteenth Street, N.W.
Washington, DC 20036

Microfilm Editions: University Microfilms International, Serial Bld coordinator, 300 North Zeeb Road, Ann Arbor, MI 48106.

The AMERICAN MATHEMATICAL MONTHLY (ISSN 0002-9890) is published monthly except bimonthly June-July and August-September by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, DC 20036 and Montpelier, VT. Copyrighted by the Mathematical Association of America (Incorporated), 1994, including rights to this journal issue as a whole and, except where otherwise noted, rights to each individual contribution. General permission is granted to Institutional Members of the MAA for noncommercial reproduction in limited quantities of individual articles (in whole or in part) provided a complete reference is made to the source. Second class postage paid at Washington, DC, and additional mailing offices. **Postmaster:** Send address changes to the American Mathematical Monthly, Membership / Subscription Department, MAA, 1529 Eighteenth Street, N.W., Washington, DC, 20036-1385.

**The American
Mathematical Monthly**

Volume 102 Number 10 / DECEMBER 1995
(ISSN 0002-9890)



Contents

ARTICLES

Polygonal Rooms Not Illuminated from Every Point /
GEORGE W. TOKARSKY 867

Three Sing-Sing Problems / GUNNAR BLOM, LARS HOLST,
and DENNIS SANDELL 880

A Nobel Prize in Mathematics / JOHN E. MORRILL 888

Some Exact Number Theory Computations via Probability Mechanisms /
RICHARD BLECKSMITH and PURUSHOTTAM W. LAUD 893

The Angle Between Complementary Subspaces / ILSE C. F. IPSEN
and CARL D. MEYER 904

FEATURES

COMMENTS 866

NOTES

The Four-Vertex Theorem Revisited—Two Variations
on the Old Theme / SERGE TABACHNIKOV 912

Entire Functions Which Vanish at Infinity / R. B. BURCKEL 916

A Converse to Cauchy's Inequality / D. ZAGIER 919

UNSOLVED PROBLEMS

Monthly Unsolved Problems, 1969–1995 / RICHARD K. GUY
and RICHARD J. NOWAKOWSKI 921

THE AUTHORS 927

PROBLEMS AND SOLUTIONS 929

REVIEWS

Modern Differential Geometry of Curves and Surfaces. By Alfred Gray /
BRUCE SOLOMON 937

TELEGRAPHIC REVIEWS 944

INDEX TO VOLUME 102 950

Polygonal Rooms Not Illuminable from Every Point

George W. Tokarsky

1. INTRODUCTION. Imagine two people in a dark room with many turns and cul-de-sacs. Assuming that the walls, floors and ceilings are constructed of reflective material, can one person strike a match and be seen by the other after repeated reflections, no matter where the two are located?

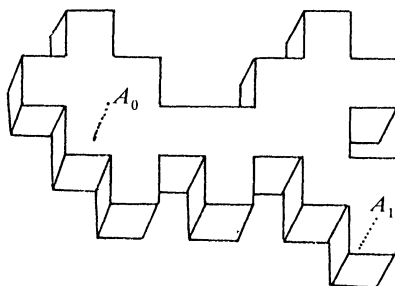


Figure 1. Can a match lit somewhere along the dotted line at A_0 say be seen somewhere along the dotted line at A_1 say?

This problem has been attributed to Ernst Straus in the early 1950's, and has remained open for over forty years. It was first published by Victor Klee in 1969 [1]. It has since reappeared on various lists of unsolved problems, notably Klee again in 1979 [2] and in two recent books on unsolved problems, one by Klee and Wagon in 1991 [3] and one by Croft, Falconer and Guy, also in 1991 [4].

In this article, we will settle the above problem in the negative. We will as well give elementary techniques for constructing rooms, both in the plane and in three-space, which are not illuminable from every point. In particular, we will show that if the two people are located in a two-dimensional planar room as shown in Figure 2, then they cannot see each other.

2. THE PLANAR PROBLEM. If G is a bounded simple polygonal region in the plane, is G illuminable from every point? In other words, if we view the sides of G as mirrors, can a single light source placed at any point, illuminate or be seen at every other point of the room? The problem can equivalently be posed in terms of a billiard ball bouncing around a pool table. Is there a "pool shot" between any two points on a polygonal pool table?

A light ray or pool ball reflects only at the sides of the room in such a way that the angle of incidence equals the angle of reflection. A light ray or pool ball that strikes a vertex is considered to end or be absorbed there.

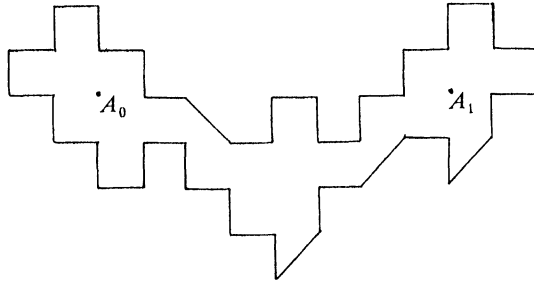


Figure 2

All paths or pool shots will be taken to be of non-zero length.

The main idea to solving this problem is that any path in a polygon unfolds to a path in another polygon constructed from mirror images of the first. Conversely, the second path can be considered to fold up to the first.

Example 1. Path $ABCD$ in 3(a) corresponds to the straight line path $ABCD$ in 3(b).

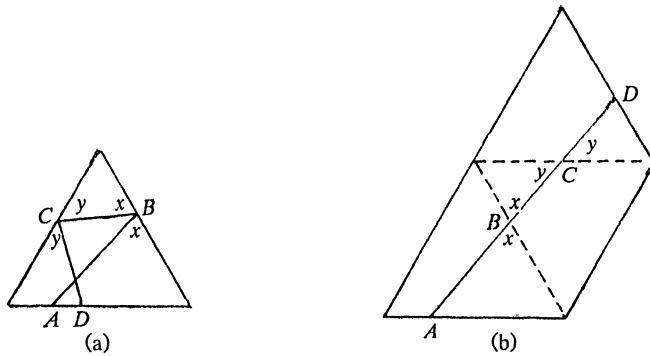


Figure 3

Example 2. Path $ABCDEF$ in 4(a) corresponds to the path $ABCDEF$ in 4(b).

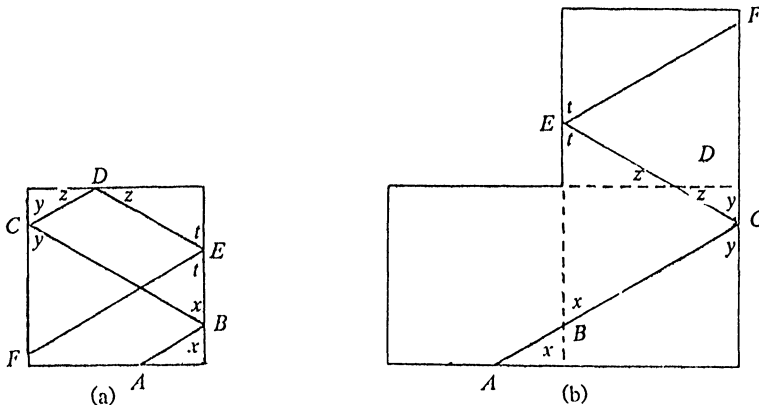


Figure 4

3. SOLUTION. We will first need the following lemma.

Lemma 3.1. *In an isosceles right triangle ABC (with right angle at C), there do not exist any pool shots from A coming back to A .*

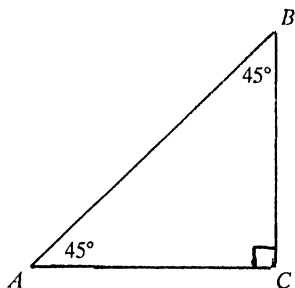


Figure 5

Proof: We start by taking a lattice of mirror images of this triangular table and assigning integer coordinates to the vertices as shown below, with A at the origin. Vertices labelled A have even coordinates $(2m, 2n)$ and vertices labelled B or C all have at least one odd coordinate. A pool shot from A to A on the original table would unfold or correspond to a straight line segment joining $A(0,0)$ to say $A(2m, 2n)$ in the lattice. This segment then must pass through the point (m, n) [or $(m/2, n/2)$ if both m and n are even, etc.] and thus must pass through a point labelled B or C . This means the pool shot would hit a vertex and be absorbed before returning to A . ■

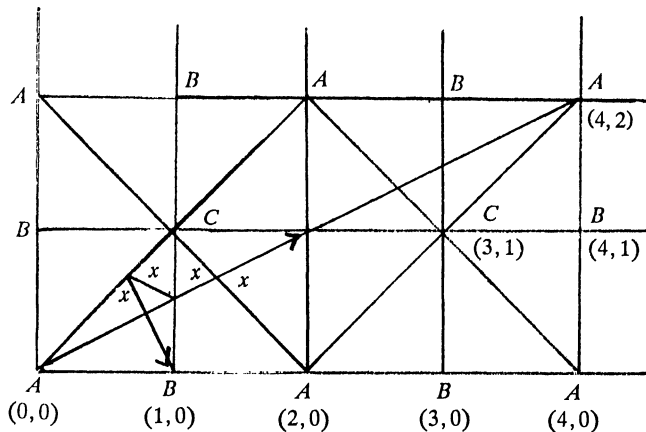


Figure 6

Theorem 3.2. *There do not exist any pool shots from A_0 to A_1 on the table shown in Figure 2.*

Proof: This table is constructed by taking mirror images of a right angled isosceles triangle as shown in Figure 7. The key to the diagram and the proof is that any point labelled B or C must be a vertex of this table, while points labelled A do not have to be.

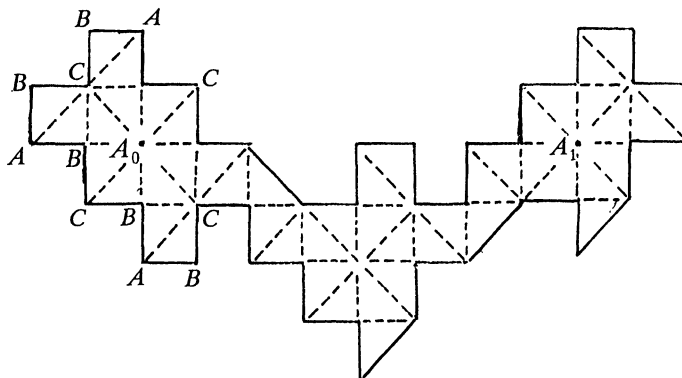


Figure 7

If there were a pool shot from A_0 to A_1 , the initial path must pass through the interior of one of the eight triangles surrounding A_0 . Let us call this triangle T . As in the lemma, a pool shot from A_0 to A_1 would correspond or fold up to a pool shot from A_0 to A_1 in triangle T , which is impossible. ■

Incidentally, it should be clear from the proof that there does not exist a pool shot between any two points labelled A on this table.

4. OTHER TABLES

Example 3. It would be interesting to find the table with the least number of sides which is not illuminable from every interior point. The example below has 26 sides.

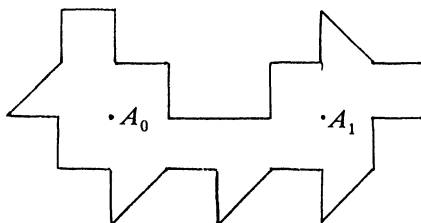


Figure 8

Example 4. By using the same kind of lattice argument given for the isosceles right triangle, there do not exist any pool shots from a corner of a square pool table $ABCD$ coming back to itself. We can also get this result by observing that a square is the mirror image of a right isosceles triangle in its hypotenuse and that a path in the square folds to a path in the triangle. A square then can be used to construct tables with only right angles, one of which is shown below. Again, we must follow the rule that points labelled B , C or D must remain vertices, while there isn't any restriction on points labelled A .

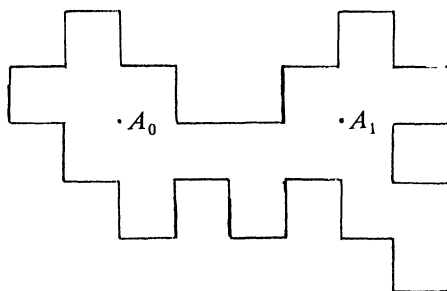


Figure 9

To construct tables using other kinds of triangles, we need a different type of argument.

Lemma 4.1. *If x divides 90 and $\angle A$ has size x° and $\angle B$ has size nx° where n is a positive integer, then the triangular pool table ABC does not have a pool shot from A coming back to A .*

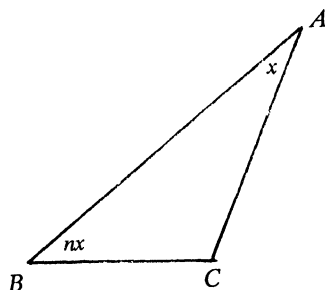


Figure 10

Proof: We measure all angles mod $2x$.

Case I. n is even.

Let $0 < \theta < x$ be the angle of a pool shot leaving A as in Figure 11(a), then inductively it bounces off sides AB and BC at angles $\pm\theta$ and side AC at angles $x \pm \theta$ as shown in Figure 11(b)(c)(d).

If it comes back to A then it must re-enter at the angle $\pm\theta \bmod 2x$, but since $0 < \theta < x$, $-\theta$ is impossible. Hence, it must re-enter at the same angle θ that it left. This can only happen if the pool shot hits one of the sides at 90° . But, then $\pm\theta \equiv 90 \bmod 2x$ which implies that $\pm\theta \equiv 0 \bmod x$ (since x divides 90) or $x \pm \theta \equiv 90 \bmod 2x$ which again implies that $\pm\theta \equiv 0 \bmod x$. This is impossible since $0 < \theta < x$.

Case II. n is odd.

Similar to the first case, a pool shot leaving A at an angle $0 < \theta < x$ hits side AB at angles $\pm\theta$, and sides BC and AC at angles $x \pm \theta$ as shown in Figure 12.

If it returns to A , then as before it must return at the same angle θ that it left. This is impossible for the same reason given above. ■

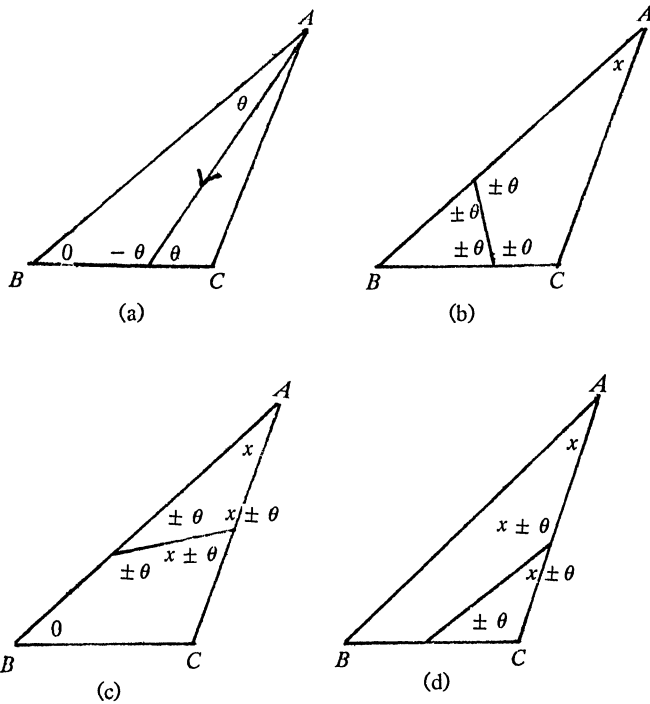


Figure 11

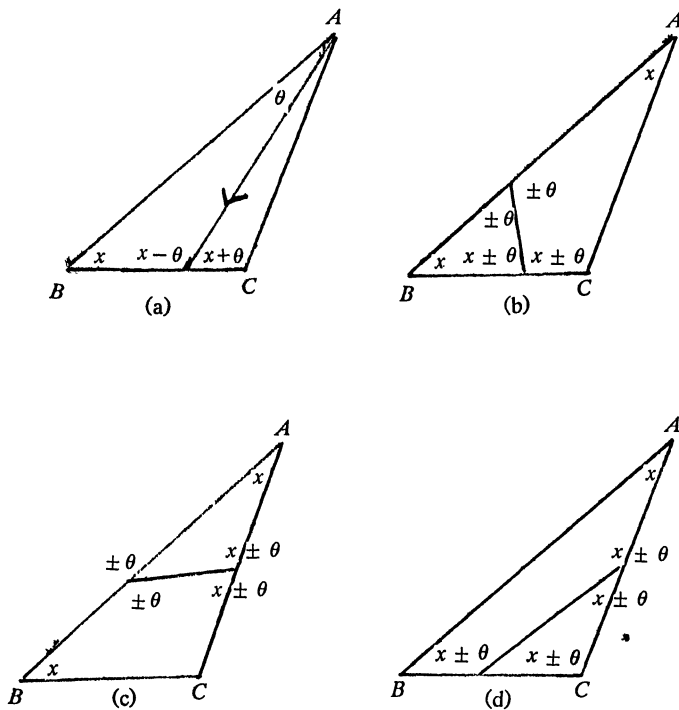


Figure 12

Example 5. On any symmetric pool table of the type shown below where x divides 90 and with angles B and C having size different from 180° (which guarantees that B and C remain vertices), there does not exist a pool shot from A to D .

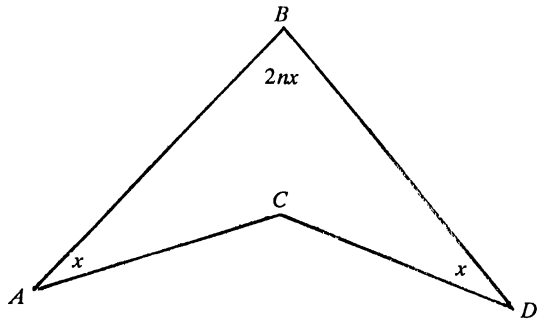


Figure 13

Proof: A pool shot from A to D would fold up to a pool shot from A to A in triangle ABC which is impossible. ■

This is an example of a quadrilateral pool table in which it is not possible to make a pool shot between two distinct boundary points.

Example 6. By the lemma, there do not exist any pool shots from A to A on the triangle ABC shown below with $m(< A) = 9^\circ$ and $n = 8$.

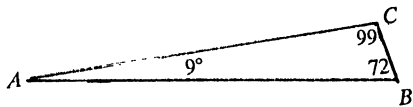


Figure 14

By taking mirror images of this triangle and following the usual rule that B and C must remain vertices, we can construct a pool table without right angles which does not have a pool shot from A_0 to A_1 .

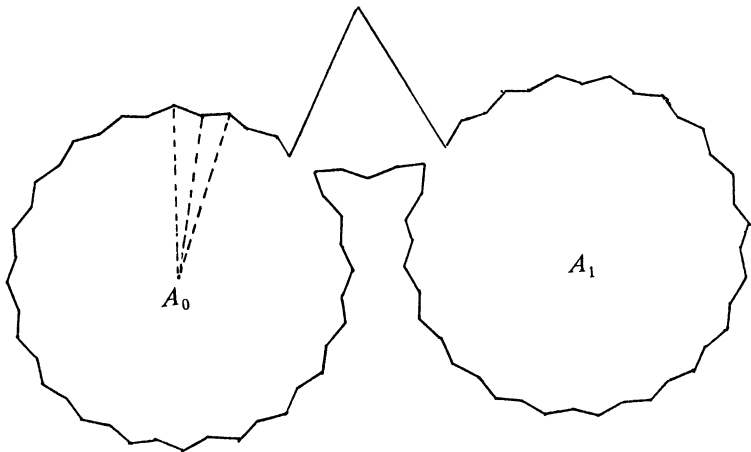


Figure 15

This example can be extended to construct pool tables with any finite number of pool shots that cannot be made.

5. GENERAL CONSTRUCTION THEOREM

Theorem 5.1. *Let G be a pool table built from a triangle ABC of the type shown in Figure 10 and which is constructed using only successive mirror images of this triangle. If G is constructed following the rule that every occurrence of B or C is a vertex, then there does not exist a pool shot between any two points labelled A .*

Proof: The pool shot is impossible by Lemma 4.1, since a path between any two points labelled A corresponds to a pool shot from A to A in triangle ABC . ■

This is the general construction result used to form the various polygonal tables.

6. THREE DIMENSIONAL EXAMPLES. In three space, reflection occurs only at points which have tangent planes, and rays bounce off the surface such that the angle between the incoming ray and the normal equals the angle between the outgoing ray and the normal. The incoming ray, the outgoing ray and the normal must be coplanar. Any ray which hits a vertex or an edge does not reflect.

If P and Q are parallel planes, it is known that a parallel projection between P and Q will preserve angles and hence reflections. This is not so if the planes are not parallel.

However, if a reflection occurs off a face whose normal \vec{n} is either perpendicular or parallel to a plane P , and Q is the plane formed by the two reflecting rays, then an orthogonal projection taking Q to P will preserve the reflection. (If \vec{n} is perpendicular to P , the projected image is a straight line).

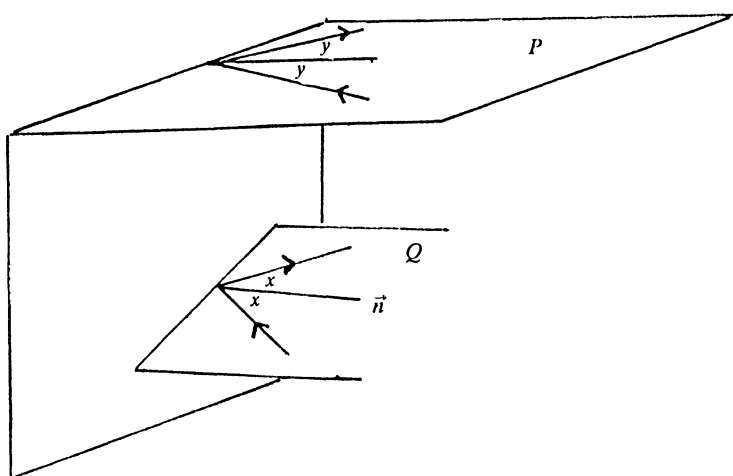


Figure 16

This means that if we form a cylinder on any of the polygonal rooms R already constructed to form a polytopal room $R \times I$, then a pool shot in the polytopal room would project orthogonally to a pool shot in R .

Example 7. The following polytopal room is not illuminable from every point. In particular there does not exist a pool shot from any point on $A_0 \times I$ to any point on $A_1 \times I$.

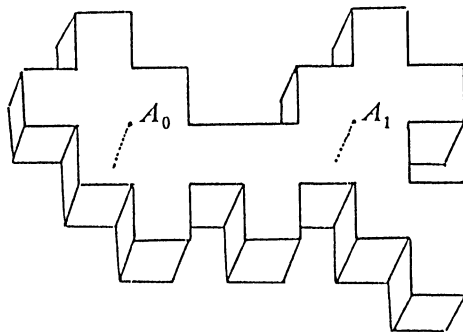


Figure 17

Proof: This would correspond to a pool shot from A_0 to A_1 in the room of Figure 9, which is impossible.

7. NON-CYLINDRICAL EXAMPLES

Lemma 7.1. *Given a cube with one corner labelled A , there do not exist any pool shots from A coming back to A .*

Proof: Let us take a lattice of mirror images of the cube with A at the origin and the vertices having integer coordinates. The A 's appear at even coordinates $(2m, 2n, 2p)$ and every other vertex has at least one odd coordinate. As before a pool shot from A to A in the original cube corresponds to a straight line segment from $A(0, 0, 0)$ to $A(2m, 2n, 2p)$ which must pass through a vertex other than A . It follows that the pool shot is impossible. ■

By virtually the same lattice argument, there does not exist a pool shot from A to any point on any edge attached to A . Alternately, we can use the projection argument with a given cube $ABCDEFGH$. If there were a reflecting path from A to X where X is on AH say, then using a suitable orthogonal projection, this path projects onto another path from A to A in the square $ABCD$ which is impossible.

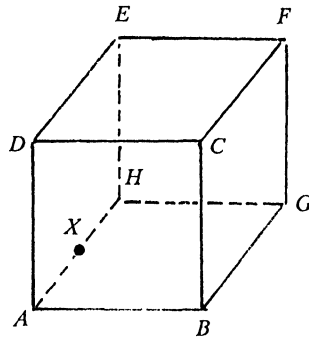
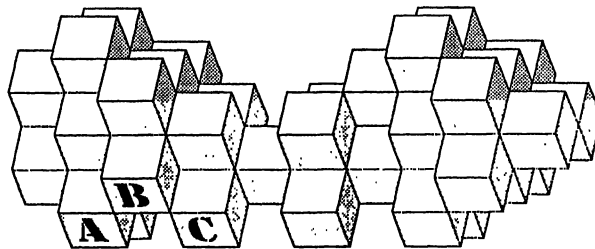


Figure 18

Example 8. It is now easy to construct a polytopal room with two interior points which are not illuminable from each other. We need only take mirror images of the cube in Figure 18 following the rule that any edge not attached to say vertex A must remain an edge. The following example was constructed in this way.



Upper View

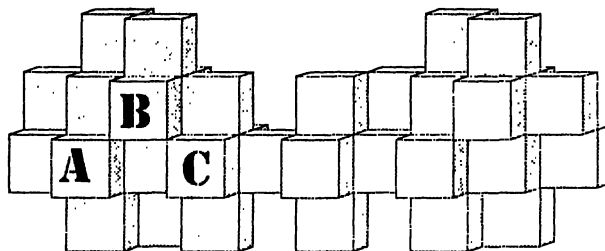


Figure 19

Proof: The above rule guarantees that a pool shot leaving A and hitting another point labelled A must pass through the interior of one of the cubes surrounding it. By the comment to Lemma 7.1 and the unfolding argument, it could never hit the second A . ■

More generally, we can use cylindrical triangular building blocks by making use of the following lemma.

Lemma 7.2. *Let T be a cylinder built on a triangle of the type shown in Figure 10, then there does not exist a pool shot between any two points on edge AD .*

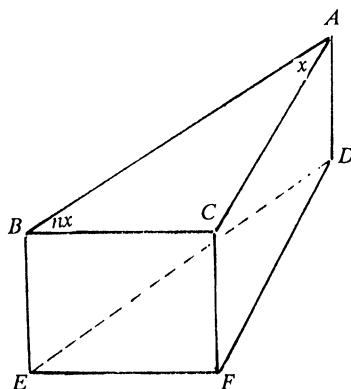


Figure 20

Proof: If we orthogonally project the cylinder onto triangle ABC , then a path in the cylinder between two points on AD corresponds to a path from A to A in triangle ABC which is impossible. ■

We immediately obtain the following result.

Three Dimensional Construction Theorem 7.3. *Let G be a polytopal room built from cylindrical triangles T of the type shown in Figure 20 and which is constructed using only successive mirror images of T . If G is constructed following the rule that every occurrence of an edge different from AD remains an edge, then*

- (a) *there does not exist a pool shot between any two points labelled A ,*
- (b) *there does not exist a pool shot between A and any D not immediately attached to A ,*
- (c) *there does not exist a pool shot between A and any interior point of a segment labelled AD which is not attached to the original A , and*
- (d) *there does not exist a pool shot between any interior point X of AD and any interior point Y of a different segment labelled AD .*

Proof: The above rule guarantees that a pool shot leaving X and hitting Y must pass through the interior of one of the cylindrical triangles surrounding X . By Lemma 7.2 and the unfolding argument, it can never hit Y . Similar proofs can be given for the other statements. ■

By symmetry, the result also holds if we interchange A and D .

8. A NON-POLYTOPAL EXAMPLE. We give a three dimensional example which is not polytopal and non-cylindrical but is a simple solid of revolution.

Example 9. If we take any symmetric quadrilateral of the type shown in Figure 13 and rotate it about the axis AD , then there does not exist a pool shot from vertex A to D , or A back to A or D back to D .

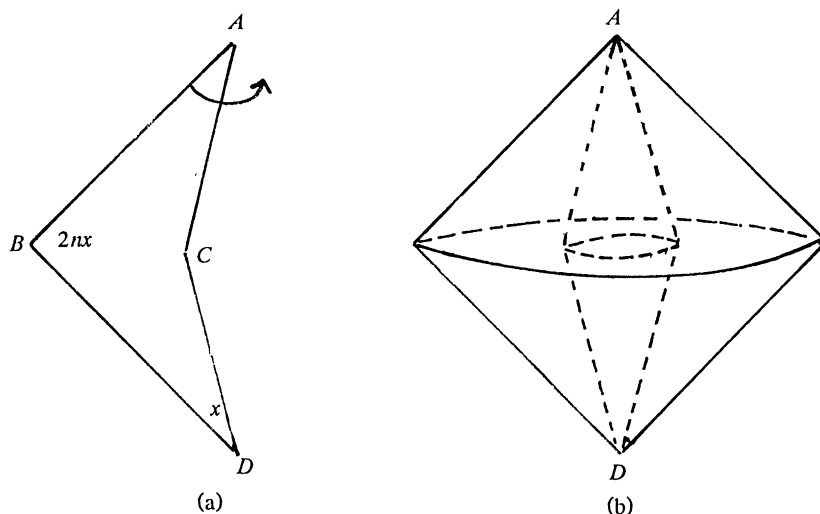


Figure 21

Proof: Rays emitted from A (say) stay in the plane determined by AD and the ray and hence correspond to a pool shot in the planar table of Figure 21(a). ■

It is possible that such a device would have physical applications, in acoustics or thermodynamics. Rays generated at A or D never reach the opposite vertex and never come back on themselves.

9. HISTORY. The illumination problem has been tentatively traced back to Ernst Straus in the early 1950's. Two questions were posed.

- (1) Is a polygonal region illuminable from every point in the region? and
- (2) Is a polygonal region illuminable from at least one point in the region?

Penrose and Penrose, 1958 [5], in an entertaining article constructed a smooth region based on properties of the ellipse which is not illuminable from various points. Other authors in written communications, then modified this example to construct a smooth region not illuminable from any point. Thus, both questions were answered negatively for smooth regions. Rauch, 1978 [6], gave an example of a smooth region not illuminable from any finite set of points.

On the other hand, the solution for polygonal regions was not forthcoming and no significant progress appeared in the literature. The nature of these problems, being easily stated and easily understood together with their apparent intractability had an obvious appeal. Thus, they started appearing on various lists of unsolved problems. Klee's paper, 1969 [1], seems to be the first published version. This was followed by a survey article of Klee and Guy, 1971 [7]. Klee again, 1979 [2], in an excellent exposition provided a list of the ten most appealing unsolved problems in plane geometry of which the illumination problem was his fifth. Recently, in 1991, two texts [3] and [4] of unsolved problems have been published both of which give excellent discussions of the two illumination problems.

I think that Klee [2] best captured the spirit of these problems in his 1979 paper subtitled, "A collection of simply stated problems that deserve equally simple solutions".

He eloquently says, "In considering the problems of this paper, it is natural to wonder whether anyone has a reasonable chance of solving them. I can't answer

that, except to say that problems of this sort are great equalizers among mathematicians, for solutions usually depend on clever ideas rather than extensive knowledge or development of complicated mathematical machinery.”

The second problem is still open.

REFERENCES

1. V. Klee, Is every polygonal region illuminable from some point? *Amer. Math. Monthly* 76 (1969), 180.
2. V. Klee, Some unsolved problems in plane geometry, *Math. Mag.* 52 (1979), 131–145.
3. V. Klee and S. Wagon, *Old and New Unsolved Problems in Plane Geometry and Number Theory*, The Math. Assoc. of America, 1991.
4. H. T. Croft, K. J. Falconer, R. K. Guy, *Unsolved Problems in Geometry*, Springer-Verlag, New York, 1991.
5. L. Penrose and R. Penrose, Puzzles for Christmas, *New Scientist*, 25 December (1958), 1580–1581, 1597.
6. J. Rauch, Illuminations of bounded domains, *Amer. Math. Monthly* 85 (1978), 359–361.
7. R. Guy and V. Klee, Monthly research problems, *Amer. Math. Monthly* 78 (1971), 1114.

Department of Mathematics
University of Alberta
632 Central Academic Building
Edmonton, Alberta, Canada
T6G 2G1

Neither you nor I nor anybody else knows what makes a mathematician tick. It is not a question of cleverness. I know many mathematicians who are far abler than I am, but they have not been so lucky. An illustration may be given by considering two miners. One may be an expert geologist, but he does not find the golden nuggets that the ignorant miner does.

—*L. J. Mordell*

Mathematical Circles Adieu. Howard W. Eves,
 Boston: Prindle, Weber and Schmidt, 1977.

Three Sing-Sing Problems

Gunnar Blom, Lars Holst and Dennis Sandell

1. INTRODUCTION AND SUMMARY. In this paper, we consider three problems, called by us the linear Sing-Sing problem, the cyclic Sing-Sing problem and the matching Sing-Sing problem. The problems are probably not new, but we have no references. However, the problems can be regarded as special matching problems in random graphs; see Barbour, Holst and Janson (1992) Section 4.4.

Linear Sing-Sing problem. If the letters in the word SINGSing are permuted at random, we may obtain, for example, INGGINSS or SNINGISG. In the former case, the permutation contains two pairs of equal neighbours, GG and SS, but in the latter case there are no such neighbours. What is the probability that the permutation has no equal neighbours? As we will show, the answer is $12/35$.

Using numbers instead of letters, the answer can be obtained from the solution to the *linear Sing-Sing problem*: What is the probability distribution of the number of equal neighbours in a random permutation of the $2n$ numbers $1122 \dots nn$?

Cyclic Sing-Sing problem. Consider the word SINGSing written in a circle:

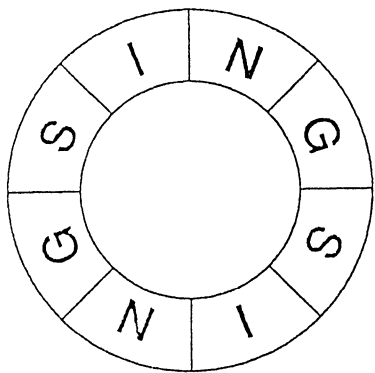


Figure 1

What is now the probability that in a random permutation there are no equal neighbours? The answer is $31/105$.

More generally, in the *cyclic Sing-Sing problem* the numbers $1122 \dots nn$ are written in a circle and we ask for the distribution of the number of equal neighbours after a random permutation.

Matching Sing-Sing problem. We write SINGSing in two rows:

S	I	N	G
S	I	N	G

and permute the letters at random. Regarding the letters in each column as neighbours, what is the probability that there are no equal neighbours? The answer is $4/7$.

As before, we generalize the situation by writing the numbers $1122 \dots nn$ in two rows:

1	2	3	...	n
1	2	3	...	n

We now obtain the *matching Sing-Sing problem*: What is the distribution of the number of equal neighbours after a random permutation?

Denote in each problem the number of equal neighbours by X_n .

In Section 2 exact expressions for the distribution and moments of X_n are obtained.

In Section 3 it is shown that, when n tends to infinity, the distribution of X_n converges to a Poisson distribution with mean 1, 1 and $1/2$ for the linear, cyclic and matching problems, respectively.

In Section 4 the accuracy of the Poisson approximations is studied by some numerical examples. Improved approximations can be obtained using the binomial distribution in the linear and the cyclic problems and the negative binomial distribution in the matching problem.

2. EXACT RESULTS

Theorem 1. *In the linear, cyclic and matching problems, the mean $E(X_n)$ is*

$$1, \frac{2n}{2n-1}, \frac{n}{2n-1},$$

respectively. The variance $\text{Var}(X_n)$ is

$$1 - \frac{1}{2n-1}, \frac{2n}{2n-1} \left(1 - \frac{1}{2n-1} \right), \frac{n}{2n-1} \left(1 + \frac{1}{(2n-1)(2n-3)} \right).$$

Note that the variance is somewhat less than the mean in the linear and the cyclic problems, but slightly larger than the mean in the matching problem. The proof of the theorem is given after the following theorem:

Theorem 2. *The probability function of X_n is given by*

$$P(X_n = k) = \sum_{j=k}^n (-1)^{j-k} \binom{j}{k} S_j$$

for $k = 0, 1, \dots, n$, where $S_0 = 1$ and

$$S_j = \binom{n}{j} \frac{2^j}{2n(2n-1) \cdots (2n-j+1)}$$

in the linear problem,

$$S_j = \frac{2n}{2n-j} \binom{n}{j} \frac{2^j}{2n(2n-1) \cdots (2n-j+1)}$$

in the cyclic problem and

$$S_j = \binom{n}{j} \frac{1}{(2n-1)(2n-3)\cdots(2n-2j+1)}$$

in the matching problem.

Proof: We first remark that S_j is the binomial moment defined by

$$E\left[\binom{X_n}{j}\right].$$

The formula for the probability function can be regarded as a generalization of the inclusion exclusion formula. It holds generally for any distribution on $0, 1, \dots, n$; see Blom, Holst and Sandell (1994, Section 3.5); the reader may also consult Feller (1968, Chapter IV). We therefore only have to prove the three given relations for the binomial moments.

Introduce in all three cases zero-one random variables I_1, \dots, I_n , where $I_i = 1$ if i and i are neighbours, and $I_i = 0$ otherwise. Then we have $X_n = I_1 + \dots + I_n$, and as the I_i 's are exchangeable, we obtain

$$E\left[\binom{X_n}{j}\right] = \binom{n}{j} E(I_1 \dots I_j) = \binom{n}{j} P(I_1 = \dots = I_j = 1).$$

It remains to calculate $P(I_1 = \dots = I_j = 1)$ in each problem.

(i) *The matching problem*

Place the numbers at random in the $2 \times n$ matrix by first putting one 1 in any entry, then the other 1 at random in any of the $2n - 1$ remaining entries, then one 2 in one of the $2n - 2$ remaining entries, etc. The probability of getting the second 1 in the same column as the first is $1/(2n - 1)$. Given that the two 1's are in the same column, the conditional probability that the 2's are in the same column is $1/(2n - 3)$. Repeating this argument we get

$$P(I_1 = \dots = I_j = 1) = \frac{1}{(2n-1)(2n-3)\cdots(2n-2j+1)},$$

which proves the assertion.

(ii) *The linear problem*

We permute $1122\dots nn$ in $n - j + 1$ steps: First, the numbers $1122\dots jj$ are ordered at random. Second, the two $(j + 1)$'s are inserted at random. Third, the two $(j + 2)$'s are inserted at random, etc.

We are interested in the event that when performing the first operation, the numbers 1 and 1, 2 and 2, \dots , and j and j become neighbours, and when performing the other operations, the pairs $11, 22, \dots, jj$ are never separated.

The probability that in the first step 1 and 1, 2 and 2, \dots , and j and j become neighbours is

$$\frac{j! 2^j}{(2j)!} = \frac{2^j}{2j(2j-1)\cdots(j+1)}.$$

When in the second step we insert $j + 1$ twice, the conditional probability that the

pairs $11, 22, \dots, jj$ are not separated is

$$\left(1 - \frac{j}{2j+1}\right) \left(1 - \frac{j}{2j+2}\right).$$

Similar conditional probabilities are obtained when $j+2$ is inserted twice, etc. Multiplying a chain of conditional probabilities, we obtain

$$\begin{aligned} P(I_1 = \dots = I_j = 1) &= \frac{2^j}{2j(2j-1) \cdots (j+1)} \prod_{k=2j+1}^{2n} \left(1 - \frac{j}{k}\right) \\ &= \frac{2^j}{2n(2n-1) \cdots (2n-j+1)}. \end{aligned}$$

(iii) *The cyclic problem*

Only a slight modification of the proof in (ii) is needed.

Proof of Theorem 1: We seek $E(X_n)$ and $\text{Var}(X_n)$. Using the same notation as in Theorem 2 we find

$$E(X_n) = E\left[\binom{X_n}{1}\right] = S_1$$

and

$$\text{Var}(X_n) = 2E\left[\binom{X_n}{2}\right] + E(X_n) - [E(X_n)]^2 = 2S_2 + S_1 - S_1^2.$$

We obtain from Theorem 2

$$S_1 = \binom{n}{1} \frac{2}{2n} = 1; \quad S_2 = \binom{n}{2} \frac{2^2}{2n(2n-1)} = \frac{n-1}{2n-1}$$

in the linear problem,

$$S_1 = \frac{2n}{2n-1} \binom{n}{1} \frac{2}{2n} = \frac{2n}{2n-1}; \quad S_2 = \frac{2n}{2n-2} \binom{n}{2} \frac{2^2}{2n(2n-1)} = \frac{n}{2n-1}$$

in the cyclic problem, and

$$\begin{aligned} S_1 &= \binom{n}{1} \frac{1}{2n-1} = \frac{n}{2n-1}; \\ S_2 &= \binom{n}{2} \frac{1}{(2n-1)(2n-3)} = \frac{n(n-1)}{2(2n-1)(2n-3)} \end{aligned}$$

in the matching problem. Introducing these three pairs of S 's in the expressions for $E(X_n)$ and $\text{Var}(X_n)$ we obtain the results in the theorem.

Remark. The original linear Sing-Sing problem is obtained by taking $n=4$ in the general linear problem in Theorem 2. This gives $S_1 = 1$, $S_2 = 3/7$, $S_3 = 2/21$ and $S_4 = 1/105$, from which we find $P(X_4 = 0) = 12/35$, as already stated at the beginning of the paper. The corresponding probability in the original cyclic problem is $31/105$ and in the original matching problem $4/7$; see Table 1.

TABLE 1. Exact probability distribution $P(X_n = k)$, $n = 2, 3, 4$, for the linear, cyclic, and matching case.

	$n = 2$			$n = 3$				$n = 4$				
k	0	1	2	0	1	2	3	0	1	2	3	4
Linear case	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{2}{5}$	$\frac{1}{5}$	$\frac{1}{15}$	$\frac{12}{35}$	$\frac{41}{105}$	$\frac{1}{5}$	$\frac{2}{35}$	$\frac{1}{105}$
Cyclic case	$\frac{1}{3}$	0	$\frac{2}{3}$	$\frac{4}{15}$	$\frac{2}{5}$	$\frac{1}{5}$	$\frac{2}{15}$	$\frac{31}{105}$	$\frac{8}{21}$	$\frac{8}{35}$	$\frac{8}{105}$	$\frac{2}{105}$
Matching case	$\frac{2}{3}$	0	$\frac{1}{3}$	$\frac{8}{15}$	$\frac{2}{5}$	0	$\frac{1}{15}$	$\frac{4}{7}$	$\frac{32}{105}$	$\frac{4}{35}$	0	$\frac{1}{105}$

3. POISSON CONVERGENCE. Let $\text{Po}(m)$ denote a Poisson distribution with mean m .

Theorem 3. *In the linear, cyclic and matching Sing-Sing problems, the number of equal neighbours X_n converges in distribution to $\text{Po}(1)$, $\text{Po}(1)$ and $\text{Po}(1/2)$, respectively, as n tends to infinity.*

Proof: From the expressions of the binomial moments of X_n in Theorem 1 it follows that these moments converge to $1/j!$, $1/j!$ and $(1/2)^j/j!$ as n tends to infinity. As these are the binomial moments of the limiting Poisson distributions, the assertion follows.

As a measure of the difference between the probability distributions of two random variables X and Y the variation distance

$$d(X, Y) = \sup_A |P(X \in A) - P(Y \in A)|$$

is sometimes used. For non-negative integer random variables it can be written in the form

$$d(X, Y) = \frac{1}{2} \sum_{k=0}^{\infty} |P(X = k) - P(Y = k)|.$$

A small value of $d(X, Y)$ indicates that it might be reasonable to approximate the distribution of X with that of Y . In the recent monograph Barbour, Holst and Janson (1992) general results on Poisson approximation are obtained and a variety of applications are studied. As a corollary of results in Section 4.4 in that book one can deduce:

Theorem 4. *Upper bounds of the variation distances between X_n and Poisson distributions with the same means as X_n in the linear, cyclic and matching problems are, respectively,*

$$\begin{aligned} & [1 - \exp(-1)] \frac{5}{2n-1}, \\ & \left[1 - \exp\left(\frac{-2n}{2n-1}\right) \right] \frac{5n-2}{n(2n-1)}, \\ & \left[1 - \exp\left(\frac{-n}{2n-1}\right) \right] \frac{4n-5}{(2n-1)(2n-3)}. \end{aligned}$$

Note that in each case the variation distance goes to zero as n goes to infinity, giving an alternative proof of Theorem 3. Also note that Theorem 4 is stronger than Theorem 3 as it also gives rates for the Poisson convergence and provides bounds on the maximal errors of the Poisson approximations.

4. APPROXIMATIONS. We shall now consider other approximations, binomial distributions in the linear and the cyclic problems and a negative binomial distribution in the matching problem. In all three cases we use distributions with the same mean and variance as X_n ; see Theorem 1. For the linear Sing-Sing problem, the distribution of X_n is approximated by a binomial distribution with parameters $2n - 1$ and $1/(2n - 1)$. In the cyclic Sing-Sing problem, the approximating binomial distribution has parameters $2n$ and $1/(2n - 1)$. The approximating distribution of X_n in the matching Sing-Sing problem is negative binomial with parameters $N = n(2n - 3)$ and $P = 1/[(2n - 1)(2n - 3)]$; here we use the following form for the probability distribution of the negative binomial:

$$p(k) = \binom{-N}{k} (-P)^k (1 + P)^{-N-k}$$

for $k = 0, 1, \dots$. Note that all three distributions are for large n close to Poisson distributions.
 In Table 2 we compare, for each of the three original Sing-Sing problems ($n = 4$), the exact probability distribution with that obtained by the Poisson

TABLE 2. Probability distribution $P(X_4 = k)$ in the second column; approximations in the third and fourth columns.

	Distribution		
k	Exact	Poisson	Alternative
Linear case			
0	0.3429	0.3679	0.3399
1	0.3905	0.3679	0.3966
2	0.2000	0.1839	0.1983
3	0.0571	0.0613	0.0551
4	0.0095	0.0153	0.0092
Cyclic case			
0	0.2952	0.3189	0.2914
1	0.3810	0.3645	0.3885
2	0.2286	0.2083	0.2266
3	0.0762	0.0793	0.0755
4	0.0190	0.0227	0.0157
Matching case			
0	0.5714	0.5647	0.5693
1	0.3048	0.3227	0.3163
2	0.1143	0.0922	0.0922
3	0.0000	0.0176	0.0188
4	0.0095	0.0025	0.0030

TABLE 3. Linear case: variation distance between true distribution and two approximating distributions.

n	Variation distance	
	Poisson	Binomial
4	0.038657	0.007063
5	0.030191	0.003967
6	0.024804	0.002559
7	0.021038	0.001787
8	0.018261	0.001319
9	0.016130	0.001014
10	0.014443	0.000805
15	0.009483	0.000337
20	0.007058	0.000184
25	0.005620	0.000116

TABLE 4. Matching case: variation distance between true distribution and two approximating distributions.

n	Variation distance	
	Poisson	Neg. Bin.
4	0.035812	0.030735
5	0.007480	0.007150
6	0.002655	0.001104
7	0.001492	0.000327
8	0.001062	0.000137
9	0.000822	0.000082
10	0.000656	0.000057
15	0.000279	0.000014
20	0.000153	0.000006
25	0.000097	0.000003

approximation and that obtained by the alternative approximation. As seen from the table the alternative approximation is very good. In Table 3 we reproduce in the linear case the variation distance between X_n and each of the two approximating distributions. Evidently, the binomial distribution provides a much better approximation than the Poisson distribution. The variation distance for the matching case is given in Table 4; clearly, the negative binomial approximation is superior to the Poisson approximation. However, to our knowledge there is no general theory available providing bounds on the error of such binomial or negative binomial approximations.

REFERENCES

1. A. D. Barbour, L. Holst, and S. Janson, *Poisson Approximation* (Oxford: Clarendon Press, 1992).
2. G. Blom, L. Holst, and D. Sandell, *Problems and Snapshots from the World of Probability* (New York: Springer-Verlag, 1994).

3. W. Feller, *An Introduction to Probability Theory and Its Applications*, Volume 1 (New York: Wiley, 1968) third edition.

Blom:
Department of Mathematical Statistics
University of Lund
Box 118
S-221 00 Lund, Sweden

Holst:
Department of Mathematics
Royal Institute of Technology
S-100 44 Stockholm, Sweden

Sandell:
Department of Biostatistics and Data Processing
Astra Draco AB
Box 34
S-221 00 Lund, Sweden

PICTURE PUZZLE
(from the collection of Paul Halmos)



How are they related?
(See page 892)

A Nobel Prize in Mathematics

John E. Morrill

On October 11, 1994, the Royal Swedish Academy of Science announced that John F. Nash was among those selected to receive a 1994 Nobel Prize. Keith Devlin, in *FOCUS*, the newsletter of the Mathematical Association of America, later wrote that this “meant for the first time in the 93-year history of the Nobel Prizes, the prize was awarded for work in pure mathematics” ([4], p. 1). But Nash had not been chosen for the award, “often referred to as the ‘Nobel Prize of Mathematics’” ([13], p. 168), the Fields Medal, or more precisely the International Medal for Outstanding Discoveries in Mathematics. Rather, he really had been selected to receive a Nobel—the Nobel Memorial Prize in Economic Science.

Upon reading this announcement, I was reminded of the many “answers” I had heard or read to the often asked question concerning “... the reason for there not being a Nobel Prize in Mathematics” (e.g., [7]). I wondered if this was really still an unsolved problem. If not, then the solution to this non-existence question is not well known, for we find in [4] the repetition of the speculation “... that a particularly bad experience in mathematics in high school led to the exclusion” More often however one hears variations on “The widely circulated explanation of why there are no Nobel Prizes in Mathematics is that Alfred Nobel wanted to make sure that the mathematician Mittag-Leffler would never be awarded a Nobel Prize” ([1], p. 39).

In my own mind the matter of the “Nobel in Mathematics” more or less had been settled by two responses to the query in that *Monthly* Letter [7]. The first, also appearing as a Letter to the Editor in this *Monthly*, [2], includes

I can offer some evidence relating to Professor King’s query about the reason for the absence of a Nobel Prize in mathematics (August–September, 1983). While I was an undergraduate at Northwestern in the early 60’s I heard the rumor that Nobel’s refusal to endow a prize in mathematics was due to a grudge against Mittag-Leffler. As I heard it then, the two had quarreled over a woman. Later I read that Mittag-Leffler had accumulated a fortune, and somehow annoyed Nobel in the process.

During a three-month visit to the Institute Mittag-Leffler in 1981 I had the opportunity to talk with archivist Barbara Bjornberg, who knows the personal lives of the people around Mittag-Leffler extremely well. She had already heard the rumor and did not believe there was any truth to it. She pointed out to me that Nobel never married and that “Mittag-Leffler’s fortune” was actually his wife’s dowry. I could not find any evidence to corroborate the rumor. Since the absence of evidence is not the same as evidence of absence, the question really must be investigated from the other end. Who first made the allegations, and on what evidence?

Such evidence as I do have suggests that there is nothing surprising in the absence of a Nobel Prize in mathematics. The question itself seems to

conceal a questionable assumption. Why *should* there be a Nobel Prize in mathematics? Nobel may not have considered mathematics important. Or, being a neighbor of Mittag-Leffler, he may have thought that the many prizes and honors Mittag-Leffler had obtained from King Oscar II for mathematicians were already sufficient recognition. These honors were quite extensive; they went to Poincaré, Appell, Bertrand, Hermite, Weierstrass, and Kovalevskaya, among others.

The second respondents, ([5], p. 73), conclude much the same.

There are Nobel prizes in physics and chemistry, so why not in mathematics? There are two current answers.

1. (French-American version) Mittag-Leffler had an affair with Nobel's wife.

2. (Swedish version) Mittag-Leffler was the leading Swedish mathematician at the time when Nobel wrote his will. Nobel knew that if there was to be a prize in mathematics, Mittag-Leffler could use his influence in the Royal Swedish Academy of Science to become the first recipient. To avoid this, Nobel gave no prize in mathematics.

Although Nobel was a confirmed bachelor, the French-American version leads a healthy life as one of the myths of mathematics and as a recurrent subject of conversation of mathematicians who think it is unfair that physics has a prize but not mathematics. The Swedish version is an academic fabrication with no credibility. In fact, Nobel and Mittag-Leffler had almost no relation to each other. The true answer to the question is that, for natural reasons, the thought of a prize in mathematics never entered Nobel's mind.

Similarly, Delvin comments "... it may simply be that Nobel felt that mathematics was not, in itself, of sufficient relevance to human development to warrant its own award" ([4], p. 1).

In looking at this story another question came to mind—has the Fields Medal ever been awarded to a (mathematical) economist? In examining this question I was led to the elegant illustrated history, *International Mathematical Congresses*, [1] and found there a citation to a work on the history of the Fields Medals. And, in this latter work, [13], I found three very interesting quotations:

"... he [Fields] spent a decade in Europe continuing his studies. "This long period of study, ... exercised a decisive influence on his life and outlook, ... Of the connections which he established, perhaps the most important was an enduring friendship with Mittag-Leffler." (p. 168)

"It was from Fields that I heard of the difficulty between Nobel and Mittag-Leffler. I gather that it was a matter of personal jealousy, ..." (p. 168)

"Perhaps I should insert here something that Fields told me and which I later verified in Sweden, namely, that Nobel hated the mathematician Mittag-Leffler and decided that mathematics would not be one of the domains in which Nobel Prizes would be available." (p. 171)

There doesn't seem to be any doubt that these last excerpts lend support to some sort of "Mittag-Leffler theory". The plot thickens. The source of all three

quotations just above is John Lighton Synge, the executor of Fields' will and a person directly involved in the establishment of the Fields Medals. The "main source" of the second respondents above—who call the Mittag-Leffler theory "an academic fabrication with no credibility"—is a book by Ragnar Sohlman, the chief executor of Nobel's will and later director of the Nobel foundation. So, if we restrict our "referees" on this matter to executors, "Why no Nobel Prize in Mathematics?" may still have the status of an unsolved problem.

On the question concerning the Fields Medal to an economist—the answer would seem to be known. No Fields Medal has gone to an economist. However, it is interesting to note that Stephen Smale, a 1966 Fields Medalist, later did do "work in economic theory" and also "joined the economics department" ([11], p. 61). It is interesting to observe the 1983 Nobel Laureate in Economic Science, Gerard Debreu, was also a Plenary Lecturer at the 1974 International Congress of Mathematicians where he spoke on "Four Aspects of the Mathematical Theory of Economic Equilibrium." Smale's comments upon the announcement of Debreu's Nobel, [11], include

Debreu's great contribution is his profound use of mathematics in the central theme of economic theory, consolidating an insight of Adam Smith more than 200 years ago. Debreu has given the foundations of general equilibrium theory in his classic work "Theory of Value." The award of the Nobel prize to Debreu gives a valuable impetus of basic research in mathematical economics.

A skimming of just the index of *Theory of Value*, [3], should convince a mathematician of the deep mathematical ideas used by some economists in their work. Of course, the mathematicalization of economics is well-known to most (and to all economics graduate students today) as is the fact that many recipients of this Nobel Prize, such as Debreu, Kenneth Arrow and Leonid Kantorovich, are also known as mathematicians. (For the interested reader descriptions of the work of Nobel Laureates in Economic Science can be found in [6] and in recent issues of *The Journal of Economic Perspectives*.)

As far as I can tell, even though Fields's endowment did not require it, no recipient of the Fields Medal has been over forty years of age. And, although not specified by the Central National Bank of Sweden who made possible the fund for the creation of the Nobel Prize in Economic Science in 1968, no Nobel Laureate in Economics has been under forty years of age. This might suggest, following from the *New York Times* headline, "Game Theory Captures a Nobel," [8], an optimal strategy for young, prize-seeking mathematicians: If you are nearing forty, and there are no prospects for a Fields shift disciplines and go for the Nobel. Since it has been claimed, ([10], p. 222),

...Simple models...in economics can exhibit dynamical behavior far more complex than anything found in classical physics or biology. In fact, all kinds of complicated dynamics (e.g., involving topological entropy, strange attractors, and even conditions yet to be found) already arise in elementary models that only describe how people exchange goods (a pure exchange model).

Instead of being an anomaly, the mathematical source of this complexity is so common to the social sciences that I suspect it highlights a general problem plaguing these areas. If true, this assertion explains why it is difficult

to achieve progress in the social sciences while underscoring the need for new mathematical tools.

Maybe one could win a Fields Medal AND a Nobel Prize, if they found the right “new mathematical tools.”

By the way, I heard a new rumor about this Nobel Prize in Mathematics thing. Fields was in Europe for a decade (1892–1902) primarily in Paris and Berlin ([12], p. 800). We already know that Fields made “an enduring friendship with Mittag-Leffler”, a man who “. . . was very conscious of the importance of maintaining a record of the contemporary history of mathematics for posterity . . .” ([9], p. 1725). Further, “In the mid-seventies Nobel settled in Paris”, but, “During the final years (1893–1896) of his life he did spend some time in Sweden . . .” ([5], p. 74).

Here’s the rumor I heard: Sometime in 1895 there was a meeting, held at Nobel’s estate Björkborn, attended by only three men, Fields, Mittag-Leffler and Nobel. There was a dispute between Fields and Nobel that apparently had begun in Paris a few years earlier, and Mittag-Leffler had been asked to arbitrate a solution. When the meeting was over and things had been decided, Nobel had gotten Literature, Physiology or Medicine, Physics, Chemistry and Peace; Fields had gotten Mathematics. As they were leaving, Fields was heard to have whispered to Mittag-Leffler “we sure put one over on him, didn’t we?”

REFERENCES

1. Donald J. Albers, G. L. Alexanderson, Constance Reid, *International Mathematical Congresses*, Springer-Verlag, 1987.
2. Roger Cooke, Letter to the Editor, *Amer. Math. Monthly* 91 (1984), 382.
3. Gerard Debreu, *Theory of Value*, John Wiley and Sons, 1959.
4. Keith Devlin, “Mathematician Awarded Nobel Prize”, *Focus* 14 (1994), 1, 5.
5. Lars Gårding and Lars Hörmander, “Why Is There No Nobel Prize in Mathematics?” *The Mathematical Intelligencer* 7 (1985), 73–74.
6. Bernard S. Katz, editor, *Nobel Laureates in Economic Sciences*, Garland, 1989.
7. Amy C. King, Letter to the Editor, *Amer. Math. Monthly* 90 (1983), 502.
8. Peter Passell, “Game Theory Captures a Nobel”, *The New York Times*, October 12, 1994, c1, c6, (National Edition).
9. Abraham Robinson, “Mittag-Leffler, Magnus Gusf (Gösta)”, *Biographical Dictionary of Mathematicians* 3, Charles Scribner’s Sons, 1991, 1724–1725.
10. Donald Saari, “Mathematical Complexity of Simple Economics,” *Notices of the American Mathematical Society* 42 (1995), 222–230.
11. Steve Smale, “Gerald Debreu Wins the Nobel Prize”, *The Mathematical Intelligencer* 6 (1984), 61–62.
12. Henry S. Tropp, “Fields, John Charles”, *Biographical Dictionary of Mathematicians* 2, Charles Scribner’s Sons, 1991, 800–801.
13. ———, “The Origins and History of the Fields Medal”, *Historia Mathematica* 3 (1976), 167–181.

Departments of Mathematics & Economics
DePauw University
5 Faculty Office Building
Greencastle, IN 46135
johnmorrill@depauw.edu

FOR WHOM NOBEL TOLLS

It is a fact that Nobel Prizes
Come in many shapes and sizes.
But one is missing from the list:
The Nobel Math Prize does not exist.

There is a widely held suspicion
Explaining this bizarre omission.
It says that *jealousy* is at the crux
Of why we get no Nobel bucks.

For Alfred Nobel became aware
Of his fiancée's prior love affair
With a *mathematician*, who held her tight
And thought that she was DYNAMITE.

Then Nobel, reacting as expected,
Vowed, "Mathematicians shall be neglected!"
"And if it's Sweden they want to see,
"Let them take a tour and pay the fee!"

Now, Nobel's behavior may be a disgrace,
Yet wouldn't the world be an even worse place
Had his lover's purported tryst
Been, instead, with a . . . *pacifist*?

—William Dunham
Department of Mathematics
Muhlenberg College
2400 Chew Street
Allentown, PA 18105

Answer to Picture Puzzle (p. 887)

The relation is that of equality:
they are both Lars Hörmander.

Some Exact Number Theory Computations via Probability Mechanisms

Richard Blecksmith and Purushottam W. Laud

1. INTRODUCTION. In this paper we apply stochastic methods to efficiently compute several number theoretic functions. The application of probability to number theory suggests density results. For example, the prime number theorem asserts that the probability that a number chosen between 1 and n is prime is approximately $1/\log n$. This is, of course, just an asymptotic estimate. Our goal here is to obtain *exact* results about functions involving the bit patterns of numbers. The results we describe can be generalized to other bases, notably base 10, but we work with base 2 for simplicity and ease of computations.

The first problem we address is how to select a random number x between 0 and a fixed bound n , written in binary as

$$n = (\epsilon_1 \epsilon_2 \cdots \epsilon_k)_2. \quad (1)$$

We assume $\epsilon_1 = 1$, so that $k = k_n$ is the number of bits in the binary expansion of n . We wish to select x by choosing its bits at random, from left to right. The problem is that we are not initially free to choose the beginning, left-most bits of x as we please. As long as the bits of the number we are forming agree with the bits of our bound n , then we are *forced* to choose 0 for the next bit of x if 0 is the corresponding bit of n , otherwise x will exceed n . Once we break the initial bit pattern of n , we may choose the next bit of x to be 0 or 1, each with probability $\frac{1}{2}$. The *bit selection chain* (or BSC) we describe is a Markov chain on the following discrete state space:

- 1: bits of x to date do not match those of n and the current bit of x is 0
- 2: bits of x to date do not match those of n and the current bit of x is 1
- 3: bits of x to date match those of n and the current bit of x is 0
- 4: bits of x to date match those of n and the current bit of x is 1.

In order to determine the transition matrix of BSC, we need the function

$$p_i = \frac{2^{k-i}}{1 + \sum_{j=i}^k \epsilon_j 2^{k-j}}. \quad (2)$$

If the i th bit of n is 1, i.e. $\epsilon_i = 1$, then p_i is the probability of the i th bit of x being 0 given that the first $i - 1$ bits of x match those of n . Observe that in the denominator of p_i , the sum $\sum_{j=i}^k \epsilon_j 2^{k-j} = (\epsilon_i \epsilon_{i+1} \cdots \epsilon_n)_2$, the number formed by truncating the first $i - 1$ leading bits of n . The transition (probability) matrix for

the i th transition is given by

$$P_i = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ p_i \epsilon_i & 0 & 1 - \epsilon_i & (1 - p_i) \epsilon_i \\ p_i \epsilon_i & 0 & 1 - \epsilon_i & (1 - p_i) \epsilon_i \end{bmatrix}.$$

The four dimensional vector $\pi^{(0)}$ is the initial distribution on the state space. After each transition, we have a distribution of the state space

$$\pi^{(i)} = \pi^{(0)} P_1 P_2 \cdots P_i = \pi^{(i-1)} P_i.$$

We denote the m th element of $\pi^{(i)}$ by $\pi_m^{(i)}$.

2. BLOCKS OF NUMBERS. We assume the binary expansion of n is given by (1). We define b_n , the number of blocks of n , to be one more than the number of i , $1 \leq i \leq k-1$, such that $\epsilon_i \neq \epsilon_{i+1}$ for n positive, $b_0 = 0$. This definition for an arbitrary base b appears in [1] where it is shown that the number of blocks of a^n written base b goes to infinity with n , as long as $\log a / \log b$ is irrational. Our goal here is to compute the sums $\sum_{j=0}^n b_j$.

We view each of the $n+1$ numbers between 0 and n , inclusive, as consisting of k bits, allowing leading bits to be 0. We have $(n+1)k$ opportunities for a switch and the number of switches is $\sum_{j=0}^n b_j$. We construct a probability mechanism that selects one of these opportunities at random (with equal probability). Then

$$Pr(\text{switch}) = \frac{1}{(n+1)k} \sum_{j=0}^n b_j.$$

The mechanism is to first select a random number by employing the bit selection chain (BSC). Then independently select a "spot" from $1, \dots, k$. Thus,

$$\begin{aligned} Pr(\text{switch}) &= \sum_{i=1}^k Pr(i\text{th spot is selected and we have a switch there}) \\ &= \sum_{i=1}^k Pr(i\text{th spot is selected}) Pr(\text{switch} \mid i\text{th spot is selected}) \\ &= \frac{1}{k} \sum_{i=1}^k Pr(\text{switch at the } i\text{th spot}). \end{aligned}$$

We choose the initial distribution to be

$$\pi^{(0)} = (0 \ 0 \ 1 \ 0).$$

Summing over the four states $1, \dots, 4$,

$$\begin{aligned} &Pr(\text{switch at the } i\text{th spot}) \\ &= \sum_{m=1}^4 Pr(\text{BSC is in state } m \text{ at } (i-1)\text{st spot}) \\ &\quad \times Pr(i\text{th transition of BSC entails a switch} \mid \text{BSC} \\ &\quad \text{is } m \text{ at the } (i-1)\text{st spot}) \\ &= \sum_{m=1}^4 \pi_m^{(i-1)} \times \text{Sum of the elements of the } m\text{th row} \\ &\quad \text{of } P_i \text{ that result in a switch.} \end{aligned}$$

Applying this to the transition matrix P_i , we find

$$\begin{aligned} \text{Pr}(\text{switch at } i\text{th spot}) &= \pi_1^{(i-1)} P_i(1, 2) + \pi_2^{(i-1)} P_i(2, 1) + \pi_3^{(i-1)} P_i(3, 4) \\ &\quad + \pi_4^{(i-1)} (P_i(4, 1) + P_i(4, 3)). \end{aligned} \quad (3)$$

Example 1. For $n = 12 = (1100)_2$, compute $\sum_{j=0}^{12} b_j$.

The four transition matrices are

$$\begin{aligned} P_1 &= \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{8}{13} & 0 & 0 & \frac{5}{13} \\ \frac{8}{13} & 0 & 0 & \frac{5}{13} \end{bmatrix} & P_2 &= \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{4}{5} & 0 & 0 & \frac{1}{5} \\ \frac{4}{5} & 0 & 0 & \frac{1}{5} \end{bmatrix} \quad \text{and} \\ P_3 &= P_4 = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}. \end{aligned}$$

Hence the probability distribution vectors $\pi^{(0)}, \dots, \pi^{(3)}$ (we don't need $\pi^{(4)}$) are: $(0, 0, 1, 0)$, $(\frac{8}{13}, 0, 0, \frac{5}{13})$, $(\frac{8}{13}, \frac{4}{13}, 0, \frac{1}{13})$, and $(\frac{6}{13}, \frac{6}{13}, \frac{1}{13}, 0)$. A routine calculation gives

$$\text{Pr}(\text{switch}) = \frac{1}{4} \sum_{i=1}^4 \text{Pr}(\text{switch at spot } i) = \frac{1}{4} \left(\frac{5}{13} + \frac{8}{13} + \frac{7}{13} + \frac{6}{13} \right) = \frac{1}{2}.$$

Thus,

$$\sum_{j=0}^{12} b_j = (n + 1) k \text{Pr}(\text{switch}) = 13 \cdot 4 \cdot \frac{1}{2} = 26,$$

which can be verified directly.

Example 2. Compute $\sum_{j=0}^n b_j$ for $n = 2^{k-1}$.

Since $p_1 = 2^{k-1}/(n + 1)$ and $1 - p_1 = 1/(n + 1)$, the first transition matrix is

$$P_1 = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{2^{k-1}}{n+1} & 0 & 0 & \frac{1}{n+1} \\ \frac{2^{k-1}}{n+1} & 0 & 0 & \frac{1}{n+1} \end{bmatrix}.$$

Since all the other bits in the binary expansion of 2^{k-1} are 0's, for $i \geq 2$ each transition matrix

$$P_i = P = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

Observe that P is an idempotent, so each power of P is just P . Thus, for $i \geq 2$, we have

$$P_1 P_2 \cdots P_i = P_1 P = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{2^{k-2}}{n+1} & \frac{2^{k-2}}{n+1} & \frac{1}{n+1} & 0 \\ \frac{2^{k-2}}{n+1} & \frac{2^{k-2}}{n+1} & \frac{1}{n+1} & 0 \end{bmatrix}.$$

The distribution vectors are the third rows of these products:

$$\begin{aligned} \pi^{(0)} &= (0 \ 0 \ 1 \ 0), \quad \pi^{(1)} = \left(\frac{2^{k-1}}{n+1} \ 0 \ 0 \ \frac{1}{n+1} \right), \\ \pi^{(i)} &= \left(\frac{2^{k-2}}{n+1} \ \frac{2^{k-2}}{n+1} \ \frac{1}{n+1} \ 0 \right) \text{ for } i \geq 2. \end{aligned}$$

Applying (3) and simplifying, we find

$$\begin{aligned} \Pr(\text{switch at spot 1}) &= \frac{1}{n+1}, \quad \Pr(\text{switch at spot 2}) = \frac{2^{k-2} + 1}{n+1}, \\ \Pr(\text{switch at spot } i) &= \frac{2^{k-2}}{n+1}, \end{aligned}$$

for $i \geq 2$. Thus, for $n = 2^{k-1}$,

$$\sum_{j=0}^n b_j = (n+1) \left[\frac{1}{n+1} + \frac{2^{k-2} + 1}{n+1} + (k-2) \frac{2^{k-2}}{n+1} \right] = (k-1)2^{k-2} + 2.$$

3. SUM OF THE RATIOS OF BLOCKS TO BITS. We denote the ratio of blocks to bits by

$$\delta_j = \frac{b_j}{k_j},$$

where k_j is the number of bits in the binary representation of j . Given a fixed integer n whose binary representation is (1), our goal is to compute the sum

$$\sum_{j=1}^n \delta_j = \sum_{j=1}^n \frac{1}{k_j} \sum_{i=1}^k S_{ij},$$

where the indicator variable S_{ij} is 1 if there is a switch at the i th bit of j , zero otherwise. To account for the denominator k_j in δ_j , we develop a probability mechanism that modifies BSC to generate integers between 1 and n with the following three conditions: (1) selecting an integer j is independent from selecting a bit i ; (2) each bit is selected with equal probability $1/k$; and (3) each integer j is selected with probability proportional to $1/k_j$. We call the constant of proportionality in the third condition $1/c$. To find c , note that since we are generating

numbers j between 1 and n ,

$$1 = \sum_{j=1}^n \Pr(\text{number selected is } j) = \sum_{j=1}^n \frac{1}{c} \frac{1}{k_j}$$

so that

$$c = \sum_{j=1}^n \frac{1}{k_j} = \sum_{i=1}^k \frac{1}{k-i+1} \cdot (\text{number of } (k-i+1) \text{ bit integers } \leq n).$$

Since $1 + \sum_{i=2}^k \epsilon_i 2^{k-i}$ enumerates the k -bit numbers between 1 and n , while 2^{k-i} enumerates the $(k-i+1)$ -bit numbers $\leq n$, we can effectively compute the normalizing constant c by the formula

$$c = \frac{1 + \sum_{i=2}^k \epsilon_i 2^{k-i}}{k} + \sum_{i=2}^k \frac{2^{k-i}}{k-i+1}. \quad (4)$$

Since S is a Bernoulli random variable, the expected value of S_{ij} is precisely the probability that $S_{ij} = 1$. It follows from our three conditions that

$$\begin{aligned} \Pr(\text{switch}) &= E(S_{ij}) = \sum_{i=1}^k \sum_{j=1}^n S_{ij} \cdot \Pr(\text{number selected is } j \text{ and bit selected is } i) \\ &= \sum_{i=1}^k \sum_{j=1}^n S_{ij} \cdot \Pr(\text{number selected is } j) \Pr(\text{bit selected is } i) \\ &= \sum_{i=1}^k \sum_{j=1}^n \frac{1}{ck_j} \frac{1}{k} S_{ij} = \frac{1}{kc} \sum_{j=1}^n \delta_j. \end{aligned}$$

Thus

$$\sum_{j=1}^n \delta_j = kc \Pr(\text{switch}).$$

Now to generate a random integer j with probability proportional to $1/k_j$, we adjust the state space of the BSC Markov chain as follows:

- 1: $\left. \begin{array}{l} 2: \\ 3: \\ 4: \end{array} \right\}$ as in BSC with the additional constraint that not all bits to date are 0
- 5: all bits to date are 0.

To compute the transition matrix for this expanded state space, note that the probability of each transition in the subchain of states 1–4 is precisely that for BSC. It is clearly impossible to move from a state 1–4 to state 5. The task remains of finding the probability of leaving state 5 at any particular stage i in the Markov chain. Let s_i denote the conditional probability that $k_j = k-i+1$ given $k_j \leq k-i+1$, that is, the probability that the i th bit is 1 given all previous bits are 0. To make sure that the marginal probability is $1/c(k-i+1)$ for the event “a specific j is generated having $k-i+1$ bits”, we must have for $i = 2, \dots, k$,

$$\Pr(\text{bits } 1, \dots, i-1 = 0, \text{ bit } i = 1) = \prod_{m=1}^{i-1} (1 - s_m) \hat{s}_i = \frac{2^{k-i}}{c(k-i+1)},$$

since there are 2^{k-i} such numbers. This gives the recursive formula

$$s_i = \frac{2^{k-i}}{c(k-i+1)\prod_{m=1}^{i-1}(1-s_m)}. \quad (5a)$$

To obtain s_1 , notice that the given event “all previous bits are 0” has probability 1 and recall that the probability a specific k bit number is generated should equal $1/ck$. Since there are $1 + \sum_{i=2}^k \epsilon_i 2^{k-i}$ such numbers between 1 and n , we have

$$s_1 = \frac{1 + \sum_{i=2}^k \epsilon_i 2^{k-i}}{ck}. \quad (5b)$$

We can now determine the transition matrix P_i for this Markov chain*

$$P_i = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ p_i \epsilon_i & 0 & 1 - \epsilon_i & (1 - p_i) \epsilon_i & 0 \\ p_i \epsilon_i & 0 & 1 - \epsilon_i & (1 - p_i) \epsilon_i & 0 \\ 0 & s_i & 0 & 0 & 1 - s_1 \end{bmatrix}.$$

The distribution vector is

$$\pi^{(j)} = \pi^{(j-1)} P_j, \quad \pi^{(1)} = (0 \ 0 \ 0 \ s_1 \ 1 - s_1).$$

By our probability mechanism,

$$Pr(\text{switch at first spot}) = s_1$$

and for $i \geq 2$,

$$\begin{aligned} Pr(\text{switch at } i\text{th spot}) &= \pi_1^{(i-1)} P_i(1, 2) + \pi_2^{(i-1)} P_i(2, 1) + \pi_3^{(i-1)} P_i(3, 4) \\ &\quad + \pi_4^{(i-1)} (P_i(4, 1) + P_i(4, 3)) + \pi_5^{(i-1)} P_i(5, 2). \end{aligned} \quad (6)$$

This is just (3) with the extra term to account for the fifth state. Finally, by the requirements that each spot is chosen with equal probability and independently of the number selected, the argument of the previous section applies to give

$$Pr(\text{switch}) = \frac{1}{k} \sum_{i=1}^k Pr(\text{switch at } i\text{th spot}).$$

Example 3. For $n = 12 = (1100)_2$, compute $\sum_{j=1}^{12} \delta_j$.

Formulas (4) and (5) give

$$c = \frac{55}{12} \quad s_1 = \frac{3}{11} \quad s_2 = \frac{3}{5} \quad s_3 = \frac{1}{2} \quad s_4 = 1.$$

*For notational simplicity, we re-use the letters P and π for the transition matrix and distribution vector, and do not worry about distinguishing them from their counterparts in the BSC chain. The appropriate usage is always clear from context.

The three transition matrices are

$$P_2 = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{8}{13} & 0 & 0 & \frac{5}{13} & 0 \\ \frac{8}{13} & 0 & 0 & \frac{5}{13} & 0 \\ 0 & \frac{2}{5} & 0 & 0 & \frac{3}{5} \end{bmatrix} \quad P_3 = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{4}{5} & 0 & 0 & \frac{1}{5} & 0 \\ \frac{4}{5} & 0 & 0 & \frac{1}{5} & 0 \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} \end{bmatrix} \quad \text{and}$$

$$P_4 = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}.$$

Hence the probability distribution vectors $\pi^{(1)}, \dots, \pi^{(3)}$ are:

$$\left(0, 0, 0, \frac{3}{11}, \frac{8}{11}\right), \quad \left(\frac{12}{55}, \frac{16}{55}, 0, \frac{3}{55}, \frac{24}{55}\right), \quad \left(\frac{14}{55}, \frac{26}{55}, \frac{3}{55}, 0, \frac{12}{55}\right).$$

A routine calculation using (6) gives

$$Pr(\text{switch}) = \frac{1}{4} \sum_{j=1}^4 Pr(\text{switch at spot } j) = \frac{1}{4} \left(\frac{3}{11} + \frac{28}{55} + \frac{29}{55} + \frac{32}{55} \right) = \frac{26}{55}.$$

Thus,

$$\sum_{j=1}^{12} \delta_j = kc Pr(\text{switch}) = 4 \cdot \frac{55}{12} \cdot \frac{26}{55} = \frac{26}{3},$$

which can be verified by a direct calculation.

4. RUDIN-SAPIRO SEQUENCE. The Rudin-Shapiro sequence $\{a_n\}$ is defined by

$$a_n = (-1)^{\sum_{i=1}^{k-1} \epsilon_i \epsilon_{i+1}}$$

where the binary representation of n is (1) . That is, the sign of a_n depends on the parity of the number of pairs of consecutive 1's in the binary expansion of n . The sequence $\{a_n\}$ relates to a variety of interesting applications from computing the orientation of creases in paper folding to spin theory related to Ising models in high energy physics. (See [4].) In this section we wish to evaluate the sum

$$s_n = \sum_{j=0}^n a_j$$

of the Rudin-Shapiro coefficients. Brillhart and Morton give formulas in [2] for calculating s_n based on doubling and shifting the index. In a subsequent paper with Erdős, they examine further analytical properties of s_n . We approach the problem of computing s_n by viewing the selection of a number j between 0 and n with Rudin-Shapiro coefficient $a_j = 1$ as a probability, which can be calculated by a Markov chain. As we form the bits of x by BSC, we keep track of the parity of the number of consecutive 1's in the bits of x to date. We include a *parity switch* in our BSC state space. This switch is 0 or 1 depending on whether the number of

pairs of successive 1's to date is even or odd. The state space for the Rudin-Shapiro Chain is

- 1: bits of x to date do not match those of n , the current bit is 0, and the current parity switch is 0
- 2: bits of x to date do not match those of n , the current bit is 1, and the current parity switch is 0
- 3: bits of x to date match those of n , the current bit is 0, and the current parity switch is 0
- 4: bits of x to date match those of n , the current bit is 1, and the current parity switch is 0
- 5: bits of x to date do not match those of n , the current bit is 0, and the current parity switch is 1
- 6: bits of x to date do not match those of n , the current bit is 1, and the current parity switch is 1
- 7: bits of x to date match those of n , the current bit is 0, and the current parity switch is 1
- 8: bits of x to date match those of n , the current bit is 1, and the current parity switch is 1.

A transition from states 1–4 to states 5–8, and vice-versa, occurs only if the current bit of both states is 1, so that the parity switch changes. Hence the transition matrix for the Rudin Shapiro sum is

$$P_i = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 \\ p_i \epsilon_i & 0 & 1 - \epsilon_i & (1 - p_i) \epsilon_i & 0 & 0 & 0 & 0 \\ p_i \epsilon_i & 0 & 1 - \epsilon_i & 0 & 0 & 0 & 0 & (1 - p_i) \epsilon_i \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & p_i \epsilon_i & 0 & 1 - \epsilon_i & (1 - p_i) \epsilon_i \\ 0 & 0 & 0 & (1 - p_i) \epsilon_i & p_i \epsilon_i & 0 & 1 - \epsilon_i & 0 \end{bmatrix}.$$

Observe that P_i has a block pattern that strongly resembles the BSC transition matrix $P_i^{(BSC)}$. If e_{ij} denotes the 4×4 matrix consisting of a 1 in position (i, j) , with all other entries 0, then $P_i = \begin{bmatrix} A_i & B_i \\ B_i & A_i \end{bmatrix}$, where $B_i = \frac{1}{2}e_{22} + (1 - p_i)\epsilon_i e_{44}$ and $A_i = P_i^{(BSC)} - B_i$. The sum of the Rudin-Shapiro coefficients s_n is given by

$$\frac{s_n}{n+1} = \pi^{(0)} P_1 P_2 \cdots P_k \sigma, \quad (7)$$

where

$$\pi^{(0)} = [0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0] \text{ and}$$

$$\sigma = [1 \ 1 \ 1 \ 1 \ -1 \ -1 \ -1 \ -1]^T.$$

The premultiplying vector $\pi^{(0)}$ identifies 3 as the initial state. The postmultiplying vector σ gives the correct sign to the final distribution, i.e. the portion ending in states 1–4 represent those j 's with $a_j = 1$, those ending in states 5–8 represent those j 's with $a_j = -1$.

Example 4. Compute $s_n = \sum_{j=0}^n a_j$ for $n = 2^{k-1}$.

For stages $2, \dots, k$, each transition matrix has the form $P_i = P = \begin{bmatrix} A & B \\ B & A \end{bmatrix}$, where

$$A = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & & \\ \frac{1}{2} & 0 & & \\ & & 1 & 0 \\ & 0 & 1 & 0 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 0 & 0 & & \\ 0 & \frac{1}{2} & & \\ & 0 & 0 & \\ & 0 & 0 & \end{bmatrix}.$$

Since the transition matrices P_i are constantly P for $i \geq 2$, (7) becomes

$$\frac{s_n}{n+1} = \pi^{(0)} P_1 P^{k-1} \sigma.$$

We wish to evaluate P^{k-1} . Although P is not an idempotent, as in Example 2, the powers of P have a fairly simple form: $P^{k-1} = \begin{bmatrix} A_{k-1} & B_{k-1} \\ B_{k-1} & A_{k-1} \end{bmatrix}$, where the blocks A_{k-1} and B_{k-1} depend on the parity of k . Define

$$d = \begin{cases} 1 & \text{if } k \text{ is odd} \\ 2 & \text{if } k \text{ is even} \end{cases} \quad \text{and} \quad e = \frac{k+d}{2}.$$

If

$$\alpha = \frac{1}{4} + \frac{1}{2^e} \quad \text{and} \quad \beta = \frac{1}{4} - \frac{1}{2^e},$$

then

$$A_{k-1} = \begin{bmatrix} \alpha & \alpha & & \\ \alpha & \beta & & \\ & & 1 & 0 \\ & 0 & 1 & 0 \end{bmatrix} \quad \text{and} \quad B_{k-1} = \begin{bmatrix} \beta & \beta & & \\ \beta & \alpha & & \\ & & 0 & 0 \\ & 0 & & 0 \end{bmatrix} \quad \text{if } k \text{ is even,}$$

$$A_{k-1} = \begin{bmatrix} \alpha & \frac{1}{4} & & \\ \frac{1}{4} & \alpha & & \\ & & 1 & 0 \\ & 0 & 1 & 0 \end{bmatrix} \quad \text{and} \quad B_{k-1} = \begin{bmatrix} \beta & \frac{1}{4} & & \\ \frac{1}{4} & \beta & & \\ & & 0 & 0 \\ & 0 & & 0 \end{bmatrix} \quad \text{if } k \text{ is odd.}$$

This formula can be proved by straightforward induction on k . The blocks for $P^k = \begin{bmatrix} A_k & B_k \\ B_k & A_k \end{bmatrix}$ are given by

$$A_k = AA_{k-1} + BB_{k-1} \quad \text{and} \quad B_k = AB_{k-1} + BA_{k-1}.$$

By examining the cases (i) k is even and (ii) k is odd, it is easy to see that A_k and B_k satisfy the above formula for the appropriate values of e , α , and β .

The product $\pi^{(0)} P_1$ is just the third row of P_1 . Since the probability used in computing the first transition matrix P_1 is $p_1 = (2^{k-1}/(n+1)) = 1 - 1/(n+1)$, we have

$$\pi^{(1)} = \pi^{(0)} P_1 = \begin{bmatrix} \frac{2^{k-1}}{n+1} & 0 & 0 & \frac{1}{n+1} & 0 & 0 & 0 & 0 \end{bmatrix}.$$

By our formula for P^{k-1} , the product $P^{k-1} \sigma$ is readily computed by subtracting the sum of columns 5–8 of P^{k-1} from the sum of columns 1–4. Combining both

cases where k is even or odd, we find

$$P^{k-1}\sigma = [d(\alpha - \beta) \quad d(\alpha - \beta) \quad 1 \quad 1 \quad d(\beta - \alpha) \quad d(\beta - \alpha) \quad -1 \quad -1]^T.$$

Thus, s_n is $(n + 1)$ times the product of these two vectors:

$$s_n = d(\alpha - \beta) \cdot 2^{k-1} + 1 = d2^{k-e} + 1,$$

since $\alpha - \beta = 2 \cdot 1/2^e$. Substituting for d and e gives

$$s_n = \begin{cases} 2^{k/2} + 1 & \text{if } k \text{ is even} \\ 2^{(k-1)/2} + 1 & \text{if } k \text{ is odd.} \end{cases}$$

5. COMPUTER IMPLEMENTATION AND CONJECTURES. We have written computer programs to implement the three methods discussed in this paper. The algorithms are extremely efficient. Consider, for example, the problem of computing the Rudin-Shapiro sum s_n . The 8×8 transition matrix P_i has either 14 or 16 nonzero entries, according as $\epsilon_i = 0$ or 1. Thus, formula (7) requires at most $16k + 8$ multiplies, where k is the number of bits of n . For the block sum and Rudin-Shapiro problem, the final answer is an integer. For the δ sum problem, however, the final answer is a rational, stored in the computer as a floating point number x . But note that the denominators in the sum are all $\leq k$. We obtain the fractional representation a/b by multiplying x by the least common multiple of the integers $1, \dots, k$.

We present a short table of the three functions for $n = 10^m$, $m = 1, \dots, 10$:

m	$\sum_{j=1}^n b_j$	$\sum_{j=1}^n \delta_j$	s_n
1.	21	$7\frac{5}{12}$	7
2.	352	$61\frac{61}{210}$	13
3.	5040	$562\frac{521}{630}$	37
4.	66918	$5420\frac{8659}{12870}$	181
5.	848344	54041.07	473
6.	10048328	530201.49	1111
7.	116578252	5223784.09	5777
8.	1350523914	52609339.74	16367
9.	15070502158	520912798.65	42209
10.	166108468238	5146376588.29	189985

The data for s_n are consistent with the result of Brillhart and Morton [2] that s_n/\sqrt{n} varies between $\sqrt{6}$ and $\sqrt{6}$ as $n \rightarrow \infty$. The data in the second column agree with the expression $\frac{1}{2}n \log_2 n$ to 3 or 4 decimal places as we read down the column. Such an asymptotic estimate is suggested by Example 2 where it is shown that $\sum_{j=0}^n b_j = \frac{1}{2}n \log_2 n + 2$ when n is a power of 2. Looking at the third column, it appears that $1/n \sum_{j=1}^n \delta_j$ is slowly approaching $\frac{1}{2}$. We generalize this to arbitrary

bases in the following

Conjecture. Given a fixed base b . Let $B(j)$ denote the number of blocks of j in its base b representation and $D(j)$ denote the number of base b digits. Then

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n \frac{B(j)}{D(j)} = \frac{b-1}{b}.$$

REFERENCES

1. R. Blecksmith, M. Filaseta, and C. Nicol, *A result on the digits of a^n* , *Acta. Arith.*, **54** (1993), 331–9.
2. J. Brillhart and P. Morton, *Über Summen von Rudin-Shapiroschen Koeffizienten*, *Illinois J. Math.*, **22** (1978), 126–148.
3. J. Brillhart, P. Erdős, and P. Morton, *On sums of the Rudin-Shapiro coefficients II*, *Pac. J. Math.*, **107** (1983), 39–69.
4. M. Mendes France and A. J. van der Poorten, *Arithmetic and analytic properties of paper folding sequences*, *Bull. Austral. Math. Soc.*, **24** (1981), 123–131.

Department of Mathematical Sciences
Northern Illinois University
DeKalb, IL 60115
richard@math.niu.edu

Division of Biostatistics
Medical College of Wisconsin
Milwaukee, WI 53226
laud@biostat.mcw.edu

Mathematics is not a deductive science—that's a cliché. When you try to prove a theorem, you don't just list the hypotheses, and then start to reason. What you do is trial and error, experimentation, guesswork.

—Paul R. Halmos

I Want to be a Mathematician. Washington DC: MAA Spectrum,
1985, p. 321.

The Angle Between Complementary Subspaces

Ilse C. F. Ipsen and Carl D. Meyer

1. INTRODUCTION. Almost all linear algebra courses discuss angles between vectors. The angle between two nonzero vectors \mathbf{u} and \mathbf{v} in \Re^n is defined as the number $0 \leq \theta \leq \pi/2$ that satisfies

$$\cos \theta = \mathbf{v}^T \mathbf{u} / \|\mathbf{v}\|_2 \|\mathbf{u}\|_2.$$

Usually the discussion stops right there, and extensions to angles between subspaces of higher dimensions are, more or less tacitly, shoved under the rug. Perhaps this is because most instructors feel that such extensions are difficult to understand, or that further effort in this direction is not worthwhile. Indeed, this makes sense for angles between general subspaces because one would have to introduce concepts like *gap* or *distance* between subspaces [7, 12], *principal (or canonical) angles* [1, 2, 15, 12], the *CS decomposition* [11, 4, 10, 6, 12], and so on. These topics are better off in a more advanced course.

However, angles between *complementary* subspaces are easier to deal with. The purpose of our article is to draw attention to some simple, though not very well known, expressions for the angle between complementary subspaces which are easily derived from the fundamental theorem of linear algebra [14] and elementary facts about matrix norms and projectors.

Angles between complementary subspaces are not just academic. They arise, for instance, in the context of controller robustness [9, 16]. Roughly speaking, the spaces associated with the controller and the plant (a system described by a set of differential equations) are complementary subspaces. The robustness of the controller is defined by the smallest perturbation that renders the system unstable, which means that the associated subspaces are no longer complementary. The system remains stable as long as perturbations are smaller than the distance between the complementary subspaces. One measure of distance is the sine of the angle between the spaces.

2. WHICH ANGLE? Before proving any theorems, we need to be precise about which angle we are talking about. As the dimension grows beyond $n > 2$, so does the wiggle room in \Re^n , and there are a host of different angles which can be defined between a pair of general subspaces. But since we wish to eventually concentrate on complementary spaces, the concept of the *minimal angle* is the most natural one to focus on.

Definition 2.1. For nonzero subspaces $\mathcal{R}, \mathcal{N} \subseteq \mathbb{R}^n$, the minimal angle between \mathcal{R} and \mathcal{N} is defined to be the number $0 \leq \theta \leq \pi/2$ that satisfies

$$\cos \theta = \max_{\substack{\mathbf{u} \in \mathcal{R}, \mathbf{v} \in \mathcal{N} \\ \|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = 1}} \mathbf{v}^T \mathbf{u}. \quad (2.1)$$

Notice that $\theta = 0$ if and only if $\mathcal{R} \cap \mathcal{N} \neq \mathbf{0}$, and $\theta = \pi/2$ if and only if $\mathcal{R} \perp \mathcal{N}$.

While (2.1) serves to define θ , it is not easy to use—especially if one wants to compute the value of θ for a given pair of subspaces. The trick in making θ more accessible is to first think in terms of projections, and then to shift the emphasis to $\sin \theta = (1 - \cos^2 \theta)^{1/2}$.

The development also requires some elementary facts concerning the standard matrix 2-norm defined by

$$\|\mathbf{A}\|_2 = \max_{\|\mathbf{x}\|_2 = 1} \|\mathbf{A}\mathbf{x}\|_2 \quad \text{for } \mathbf{A} \in \mathbb{R}^{n \times n} \quad \text{and} \quad \mathbf{x} \in \mathbb{R}^{n \times 1}.$$

The following properties can be found (often as exercises) in standard texts.

$$\|\mathbf{A}^T\|_2 = \|\mathbf{A}\|_2 \quad (2.2)$$

$$\|\mathbf{X}\mathbf{A}\mathbf{Y}\|_2 = \|\mathbf{A}\|_2 \quad \text{when } \mathbf{X} \text{ has orthonormal columns and } \mathbf{Y} \text{ has orthonormal rows} \quad (2.3)$$

$$\|\mathbf{A}\|_2 = \max_{\substack{\|\mathbf{x}\|_2 \leq 1 \\ \|\mathbf{y}\|_2 \leq 1}} \mathbf{y}^T \mathbf{A} \mathbf{x} \quad (2.4)$$

$$\|\mathbf{A}\|_2 = \frac{1}{\min_{\|\mathbf{x}\|_2 = 1} \|\mathbf{A}^{-1} \mathbf{x}\|_2} \quad \text{when } \mathbf{A}^{-1} \text{ exists} \quad (2.5)$$

$$\left\| \begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{pmatrix} \right\|_2 = \max\{\|\mathbf{A}\|_2, \|\mathbf{B}\|_2\}. \quad (2.6)$$

The first step in unraveling (2.1) is to express $\cos \theta$ in terms of the orthogonal projectors onto \mathcal{R} and \mathcal{N} .

Theorem 2.1. If $\mathbf{P}_{\mathcal{R}}$ and $\mathbf{P}_{\mathcal{N}}$ are the orthogonal projectors onto \mathcal{R} and \mathcal{N} , respectively, then

$$\cos \theta = \|\mathbf{P}_{\mathcal{N}} \mathbf{P}_{\mathcal{R}}\|_2 = \|\mathbf{P}_{\mathcal{R}} \mathbf{P}_{\mathcal{N}}\|_2. \quad (2.7)$$

Proof: For vectors \mathbf{x} and \mathbf{y} such that $\|\mathbf{x}\|_2 = \|\mathbf{y}\|_2 = 1$, we have $\mathbf{P}_{\mathcal{R}} \mathbf{x} \in \mathcal{R}$ and $\mathbf{P}_{\mathcal{N}} \mathbf{y} \in \mathcal{N}$ where $\|\mathbf{P}_{\mathcal{R}} \mathbf{x}\|_2 \leq \|\mathbf{P}_{\mathcal{R}}\|_2 \|\mathbf{x}\|_2 \leq 1$ and $\|\mathbf{P}_{\mathcal{N}} \mathbf{y}\|_2 \leq \|\mathbf{P}_{\mathcal{N}}\|_2 \|\mathbf{y}\|_2 \leq 1$, so that (2.4) can be used to write

$$\cos \theta = \max_{\substack{\mathbf{u} \in \mathcal{R}, \mathbf{v} \in \mathcal{N} \\ \|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = 1}} \mathbf{v}^T \mathbf{u} = \max_{\substack{\mathbf{u} \in \mathcal{R}, \mathbf{v} \in \mathcal{N} \\ \|\mathbf{u}\|_2 \leq 1, \|\mathbf{v}\|_2 \leq 1}} \mathbf{v}^T \mathbf{u} = \max_{\substack{\|\mathbf{x}\|_2 \leq 1 \\ \|\mathbf{y}\|_2 \leq 1}} \mathbf{y}^T \mathbf{P}_{\mathcal{N}} \mathbf{P}_{\mathcal{R}} \mathbf{x} = \|\mathbf{P}_{\mathcal{N}} \mathbf{P}_{\mathcal{R}}\|_2.$$

The fact that $\|\mathbf{P}_{\mathcal{N}} \mathbf{P}_{\mathcal{R}}\|_2 = \|\mathbf{P}_{\mathcal{R}} \mathbf{P}_{\mathcal{N}}\|_2$ is a consequence of the symmetry of orthogonal projectors together with (2.2). ■

Theorem 2.1 does not depend on \mathcal{R} and \mathcal{N} being complementary subspaces—it is a statement about the minimal angle between any two subspaces of \mathbb{R}^n . But in the special case when \mathcal{R} and \mathcal{N} are complementary, there is a more natural projector which gives rise to a formula which is simpler than (2.7).

3. ENTER THE OBLIQUE PROJECTOR.

Definition 3.1. Subspaces $\mathcal{R}, \mathcal{N} \subseteq \mathfrak{R}^n$ are said to be complementary whenever $\mathcal{R} + \mathcal{N} = \mathfrak{R}^n$ and $\mathcal{R} \cap \mathcal{N} = \mathbf{0}$, and this is denoted by writing $\mathcal{R} \oplus \mathcal{N} = \mathfrak{R}^n$. The associated oblique projector is the unique idempotent matrix \mathbf{P} whose range is \mathcal{R} and whose nullspace is \mathcal{N} . As an operator, \mathbf{P} projects vectors in \mathfrak{R}^n onto \mathcal{R} along (or parallel to) \mathcal{N} , and thus it acts as the identity on \mathcal{R} and the zero operator on \mathcal{N} .

The goal is to simplify (2.7) in the case of complementary spaces by somehow using the more natural oblique projector \mathbf{P} instead of the two orthogonal projectors $\mathbf{P}_{\mathcal{R}}$ and $\mathbf{P}_{\mathcal{N}}$. But to realize a simplification, we must shift the emphasis to $\sin \theta$ rather than $\cos \theta$.

Theorem 3.1. Suppose that $\mathcal{R}, \mathcal{N} \subset \mathfrak{R}^n$ are nonzero complementary spaces, and let \mathbf{P} be the oblique projector onto \mathcal{R} along \mathcal{N} . The minimal angle θ between \mathcal{R} and \mathcal{N} satisfies

$$\sin \theta = \frac{1}{\|\mathbf{P}\|_2}. \quad (3.1)$$

Proof: Decompose \mathbf{P} in terms of its four fundamental subspaces by choosing orthogonal matrices $\mathbf{U} = (\mathbf{U}_1 \mid \mathbf{U}_2)$ and $\mathbf{V} = (\mathbf{V}_1 \mid \mathbf{V}_2)$ in which the columns of \mathbf{U}_1 and \mathbf{U}_2 constitute orthonormal bases for \mathcal{R} and \mathcal{R}^\perp , respectively, and \mathbf{V}_1 and \mathbf{V}_2 are orthonormal bases for \mathcal{N}^\perp and \mathcal{N} , respectively, so that $\mathbf{U}_i^T \mathbf{U}_i = \mathbf{I}$ and $\mathbf{V}_i^T \mathbf{V}_i = \mathbf{I}$ for $i = 1, 2$, and

$$\mathbf{P}_{\mathcal{R}} = \mathbf{U}_1 \mathbf{U}_1^T, \quad \mathbf{I} - \mathbf{P}_{\mathcal{R}} = \mathbf{U}_2 \mathbf{U}_2^T, \quad \mathbf{P}_{\mathcal{N}} = \mathbf{V}_2 \mathbf{V}_2^T, \quad \mathbf{I} - \mathbf{P}_{\mathcal{N}} = \mathbf{V}_1 \mathbf{V}_1^T.$$

The matrices \mathbf{U} and \mathbf{V} decompose \mathbf{P} in the sense that

$$\mathbf{U}^T \mathbf{P} \mathbf{V} = \begin{pmatrix} \mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \quad \text{or, equivalently, } \mathbf{P} = \mathbf{U} \begin{pmatrix} \mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{V}^T = \mathbf{U}_1 \mathbf{C} \mathbf{V}_1^T \quad (3.2)$$

in which $\mathbf{C} = \mathbf{U}_1^T \mathbf{P} \mathbf{V}_1$ is nonsingular. (For instance, one can choose \mathbf{U} and \mathbf{V} so that this is the singular value decomposition of \mathbf{P} .) Notice that $\mathbf{P}^2 = \mathbf{P}$ implies $\mathbf{C} = \mathbf{C} \mathbf{V}_1^T \mathbf{U}_1 \mathbf{C}$, which in turn insures $\mathbf{C}^{-1} = \mathbf{V}_1^T \mathbf{U}_1$. Consequently, (2.3) together with (2.5) implies that

$$\|\mathbf{P}\|_2 = \|\mathbf{C}\|_2 = \frac{1}{\min_{\|\mathbf{x}\|_2=1} \|\mathbf{C}^{-1} \mathbf{x}\|_2} = \frac{1}{\min_{\|\mathbf{x}\|_2=1} \|\mathbf{V}_1^T \mathbf{U}_1 \mathbf{x}\|_2}.$$

Combining this with the result of Theorem 2.1 produces

$$\begin{aligned} \sin^2 \theta &= 1 - \cos^2 \theta = 1 - \|\mathbf{P}_{\mathcal{N}} \mathbf{P}_{\mathcal{R}}\|_2^2 = 1 - \|\mathbf{V}_2 \mathbf{V}_2^T \mathbf{U}_1 \mathbf{U}_1^T\|_2^2 \\ &= 1 - \|(\mathbf{I} - \mathbf{V}_1 \mathbf{V}_1^T) \mathbf{U}_1\|_2^2 = 1 - \max_{\|\mathbf{x}\|_2=1} \|(\mathbf{I} - \mathbf{V}_1 \mathbf{V}_1^T) \mathbf{U}_1 \mathbf{x}\|_2^2 \\ &= 1 - \max_{\|\mathbf{x}\|_2=1} \mathbf{x}^T \mathbf{U}_1^T (\mathbf{I} - \mathbf{V}_1 \mathbf{V}_1^T) \mathbf{U}_1 \mathbf{x} = 1 - \max_{\|\mathbf{x}\|_2=1} (1 - \|\mathbf{V}_1^T \mathbf{U}_1 \mathbf{x}\|_2^2) \\ &= 1 - \left(1 - \min_{\|\mathbf{x}\|_2=1} \|\mathbf{V}_1^T \mathbf{U}_1 \mathbf{x}\|_2^2\right) \\ &= \frac{1}{\|\mathbf{P}\|_2^2}. \quad \blacksquare \end{aligned}$$

The expression $\sin \theta = 1/\|\mathbf{P}\|_2$ is not only conceptually simple, but, as illustrated in Figure 1, there is also a particularly nice picture that accompanies it. The image of the unit sphere in \mathbb{R}^3 under \mathbf{P} is obtained by projecting all vectors on the sphere onto \mathcal{R} along lines parallel to \mathcal{N} . The result is an ellipse in \mathcal{R} .

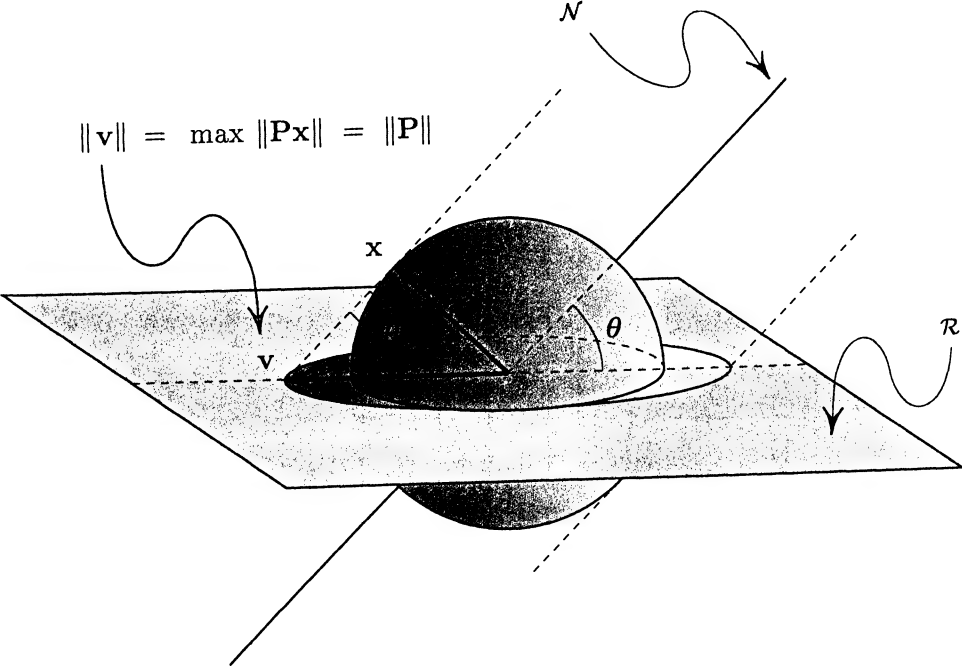


Figure 1

The norm of a longest vector \mathbf{v} on this ellipse equals the norm of \mathbf{P} , i.e.

$$\|\mathbf{v}\|_2 = \max_{\|\mathbf{x}\|_2=1} \|\mathbf{P}\mathbf{x}\|_2 = \|\mathbf{P}\|_2.$$

It is apparent from the right triangle in Figure 1 that

$$\sin \theta = \frac{\|\mathbf{x}\|_2}{\|\mathbf{v}\|_2} = \frac{1}{\|\mathbf{v}\|_2} = \frac{1}{\|\mathbf{P}\|_2}.$$

4. BACK TO ORTHOGONAL PROJECTORS. For subspaces $\mathcal{R}, \mathcal{N} \subseteq \mathbb{R}^n$ such that $\dim \mathcal{R} = \dim \mathcal{N}$, the difference $\mathbf{P}_{\mathcal{R}} - \mathbf{P}_{\mathcal{N}}$ of the associated orthogonal projectors is of special interest because $\|\mathbf{P}_{\mathcal{R}} - \mathbf{P}_{\mathcal{N}}\|_2$ is a common measure of the distance or separation between \mathcal{R} and \mathcal{N} . It is therefore natural to inquire about what can be said about the minimal angle between complementary spaces in terms of the difference $\mathbf{P}_{\mathcal{R}} - \mathbf{P}_{\mathcal{N}}$. The following theorem provides some answers.

Theorem 4.1. For nonzero subspaces $\mathcal{R}, \mathcal{N} \subseteq \mathbb{R}^n$, let $\mathbf{P}_{\mathcal{R}}$ and $\mathbf{P}_{\mathcal{N}}$ denote the orthogonal projectors onto \mathcal{R} and \mathcal{N} , respectively, and let θ be the minimal angle between \mathcal{R} and \mathcal{N} . The following two statements are true.

- \mathcal{R} and \mathcal{N} are complementary spaces if and only if $\mathbf{P}_{\mathcal{R}} - \mathbf{P}_{\mathcal{N}}$ is nonsingular. (4.1)
- If \mathcal{R} and \mathcal{N} are complementary spaces, then $\sin \theta = 1/\|(\mathbf{P}_{\mathcal{R}} - \mathbf{P}_{\mathcal{N}})^{-1}\|_2$. (4.2)

Proof of (4.1): The orthogonal matrices \mathbf{U} and \mathbf{V} which were introduced in the proof of Theorem 3.1 to decompose \mathbf{P} also decompose $\mathbf{P}_{\mathcal{R}} - \mathbf{P}_{\mathcal{N}}$ in the sense that

$$\begin{aligned} \mathbf{U}^T(\mathbf{P}_{\mathcal{R}} - \mathbf{P}_{\mathcal{N}})\mathbf{V} &= \begin{pmatrix} \mathbf{U}_1^T \\ \mathbf{U}_2^T \end{pmatrix} (\mathbf{U}_1\mathbf{U}_1^T - \mathbf{V}_2\mathbf{V}_2^T)(\mathbf{V}_1 | \mathbf{V}_2) \\ &= \begin{pmatrix} \mathbf{U}_1^T \mathbf{V}_1 & \mathbf{0} \\ \mathbf{0} & -\mathbf{U}_2^T \mathbf{V}_2 \end{pmatrix}. \end{aligned} \quad (4.3)$$

Assume first that \mathcal{R} and \mathcal{N} are nonzero complementary subspaces. If $\dim \mathcal{R} = r$, then $\mathbf{U}_1^T \mathbf{V}_1$ is $r \times r$ and $\mathbf{U}_2^T \mathbf{V}_2$ is $n - r \times n - r$, so $\mathbf{P}_{\mathcal{R}} - \mathbf{P}_{\mathcal{N}}$ is nonsingular if and only if $\mathbf{U}_1^T \mathbf{V}_1$ and $\mathbf{U}_2^T \mathbf{V}_2$ are each nonsingular. But we already know from the proof of Theorem 3.1 that $\mathbf{U}_1^T \mathbf{V}_1 = (\mathbf{C}^{-1})^T$ is nonsingular, so we only need to prove that $\mathbf{U}_2^T \mathbf{V}_2$ is nonsingular. If \mathbf{P} is the oblique projector onto \mathcal{R} along \mathcal{N} , then

$$\mathbf{P}\mathbf{U}_1 = \mathbf{U}_1 \quad \text{and} \quad \mathbf{P}\mathbf{V}_2 = \mathbf{0},$$

so that

$$\mathbf{V}_2^T(\mathbf{I} - \mathbf{P})\mathbf{U}_2\mathbf{U}_2^T\mathbf{V}_2 = \mathbf{V}_2^T(\mathbf{I} - \mathbf{P})(\mathbf{I} - \mathbf{U}_1\mathbf{U}_1^T)\mathbf{V}_2 = \mathbf{V}_1^T\mathbf{V}_1 = \mathbf{I}.$$

Thus $\mathbf{U}_2^T \mathbf{V}_2$ is nonsingular with $(\mathbf{U}_2^T \mathbf{V}_2)^{-1} = \mathbf{V}_2^T(\mathbf{I} - \mathbf{P})\mathbf{U}_2$, and consequently $\mathbf{P}_{\mathcal{R}} - \mathbf{P}_{\mathcal{N}}$ is nonsingular. Conversely, if $\mathbf{P}_{\mathcal{R}} - \mathbf{P}_{\mathcal{N}}$ is nonsingular, and if $\dim \mathcal{R} = r > 0$ and $\dim \mathcal{N} = k > 0$, then $\mathbf{U}_1^T \mathbf{V}_1$ is $r \times n - k$ and $\mathbf{U}_2^T \mathbf{V}_2$ is $n - r \times k$, so (4.3) insures that the rows as well as the columns in each of these products must be linearly independent. In other words, $\mathbf{U}_1^T \mathbf{V}_1$ and $\mathbf{U}_2^T \mathbf{V}_2$ must both be square and nonsingular, so $k = n - r$. Let

$$\mathbf{Q} = \mathbf{U}_1(\mathbf{V}_1^T \mathbf{U}_1)^{-1} \mathbf{V}_1^T,$$

and notice that $\mathbf{Q} = \mathbf{Q}^2$, so that \mathbf{Q} is a projector. If $R(*)$ and $N(*)$ denote range and nullspace, respectively, then

$$R(\mathbf{Q}) \subseteq R(\mathbf{U}_1) = R(\mathbf{Q}\mathbf{U}_1) \subseteq R(\mathbf{Q}) \Rightarrow R(\mathbf{Q}) = \mathcal{R},$$

and

$$\left\{ \begin{array}{l} N(\mathbf{Q}) \supseteq N(\mathbf{V}_1^T) = \mathcal{N} \\ \dim N(\mathbf{Q}) = n - \dim R(\mathbf{Q}) = n - r = k = \dim \mathcal{N} \end{array} \right\} \Rightarrow N(\mathbf{Q}) = \mathcal{N}.$$

In other words, $\mathbf{Q} = \mathbf{P}$ is the oblique projector onto \mathcal{R} along \mathcal{N} . Therefore, since the range and nullspace of any projector are complementary spaces, it must be the case that $\mathcal{R} \oplus \mathcal{N} = \mathfrak{R}^n$. ■

Proof of (4.2): If \mathcal{R} and \mathcal{N} are complementary, then $\mathbf{P}_{\mathcal{R}} - \mathbf{P}_{\mathcal{N}}$ is nonsingular, and (4.3) together with (2.6) can be used to conclude that

$$\|(\mathbf{P}_{\mathcal{R}} - \mathbf{P}_{\mathcal{N}})^{-1}\|_2 = \max\left\{\|(\mathbf{U}_1^T \mathbf{V}_1)^{-1}\|_2, \|(\mathbf{U}_2^T \mathbf{V}_2)^{-1}\|_2\right\}. \quad (4.4)$$

But $\|(\mathbf{U}_1^T \mathbf{V}_1)^{-1}\|_2 = \|(\mathbf{U}_2^T \mathbf{V}_2)^{-1}\|_2$ because we can again use (2.5) to write

$$\begin{aligned} \frac{1}{\|(\mathbf{U}_1^T \mathbf{V}_1)^{-1}\|_2^2} &= \min_{\|\mathbf{x}\|_2=1} \|\mathbf{U}_1^T \mathbf{V}_1 \mathbf{x}\|_2^2 = \min_{\|\mathbf{x}\|_2=1} \mathbf{x}^T \mathbf{V}_1^T \mathbf{U}_1 \mathbf{U}_1^T \mathbf{V}_1 \mathbf{x} \\ &= \min_{\|\mathbf{x}\|_2=1} \mathbf{x}^T \mathbf{V}_1^T (\mathbf{I} - \mathbf{U}_2 \mathbf{U}_2^T) \mathbf{V}_1 \mathbf{x} \\ &= \min_{\|\mathbf{x}\|_2=1} (1 - \mathbf{x}^T \mathbf{V}_1^T \mathbf{U}_2 \mathbf{U}_2^T \mathbf{V}_1 \mathbf{x}) \\ &= 1 - \max_{\|\mathbf{x}\|_2=1} \|\mathbf{U}_2^T \mathbf{V}_1 \mathbf{x}\|_2^2 = 1 - \|\mathbf{U}_2^T \mathbf{V}_1\|_2^2, \end{aligned}$$

and a similar argument proves that

$$\frac{1}{\|(\mathbf{U}_2^T \mathbf{V}_2)^{-1}\|_2^2} = 1 - \|\mathbf{U}_2^T \mathbf{V}_1\|_2^2.$$

Therefore, the results of Theorem 3.1 insure that

$$\|(\mathbf{P}_{\mathcal{R}} - \mathbf{P}_{\mathcal{N}})^{-1}\|_2 = \|(\mathbf{U}_1^T \mathbf{V}_1)^{-1}\|_2 = \|\mathbf{C}^T\|_2 = \|\mathbf{C}\|_2 = \|\mathbf{P}\|_2 = \frac{1}{\sin \theta}. \quad \blacksquare$$

Theorem 3.1 is not new—Gohberg and Kreĭn [5] attribute it to Ljance [8]—but it seems to have escaped the notice of many writers and teachers of linear algebra. We have not seen Theorem 4.1 in the literature.

5. CONSEQUENCES. Although the following facts about projectors are often proved by separate (and sometimes substantial) arguments, they turn out to be immediate consequences of Theorem 3.1 and Theorem 4.1.

Corollary 5.1. $\|\mathbf{P}\|_2 \geq 1$ for every non-zero projector \mathbf{P} , and $\|\mathbf{P}\|_2 = 1$ if and only if \mathbf{P} is an orthogonal projector.

Corollary 5.2. $\|\mathbf{I} - \mathbf{P}\|_2 = \|\mathbf{P}\|_2$ for all projectors \mathbf{P} that are not zero and not equal to the identity.

Corollary 5.3. Let \mathbf{u} and \mathbf{v} be vectors in \Re^n with $\mathbf{v}^T \mathbf{u} = 1$. If θ is the minimal angle between \mathbf{u} and \mathbf{v}^\perp (the space orthogonal to \mathbf{v}), then

$$\|\mathbf{I} - \mathbf{u}\mathbf{v}^T\|_2 = \|\mathbf{u}\mathbf{v}^T\|_2 = \frac{1}{\sin \theta} = \|\mathbf{u}\|_2 \|\mathbf{v}\|_2.$$

Proof: The first equality follows from Corollary 5.2 and the second one from Theorem 3.1. The fact that $\|\mathbf{u}\mathbf{v}^T\|_2 = \|\mathbf{u}\|_2 \|\mathbf{v}\|_2$ follows from properties of the two-norm because

$$\|\mathbf{u}\|_2 \|\mathbf{v}\|_2 = \frac{\|\mathbf{u}\mathbf{v}^T \mathbf{v}\|_2}{\|\mathbf{v}\|_2} \leq \max_{\mathbf{x} \neq 0} \frac{\|\mathbf{u}\mathbf{v}^T \mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \|\mathbf{u}\mathbf{v}^T\|_2 \leq \|\mathbf{u}\|_2 \|\mathbf{v}\|_2. \quad \blacksquare$$

Corollary 5.4. If θ is the minimal angle between complementary spaces $\mathcal{R}, \mathcal{N} \subset \Re^n$, and if θ^\perp is the minimal angle between \mathcal{R}^\perp and \mathcal{N}^\perp , then $\theta = \theta^\perp$.

Proof: This follows from Theorem 3.1 together with Corollary 5.2. The result is also a corollary of Theorem 4.1 because

$$\|(\mathbf{P}_{\mathcal{R}^\perp} - \mathbf{P}_{\mathcal{N}^\perp})^{-1}\|_2 = \|((\mathbf{I} - \mathbf{P}_{\mathcal{R}}) - (\mathbf{I} - \mathbf{P}_{\mathcal{N}}))^{-1}\|_2 = \|(\mathbf{P}_{\mathcal{R}} - \mathbf{P}_{\mathcal{N}})^{-1}\|_2.$$

Corollary 5.5. For complementary spaces $\mathcal{R}, \mathcal{N} \subset \Re^n$, let \mathbf{P} be the oblique projector onto \mathcal{R} along \mathcal{N} , and let \mathbf{Q} denote the oblique projector onto \mathcal{R}^\perp along \mathcal{N}^\perp . If $\mathbf{P}_{\mathcal{R}}$ and $\mathbf{P}_{\mathcal{N}}$ are the orthogonal projectors onto \mathcal{R} and \mathcal{N} , respectively, and if θ is the minimal angle between \mathcal{R} and \mathcal{N} , then each of the following statements is true.

- $(\mathbf{P}_{\mathcal{R}} - \mathbf{P}_{\mathcal{N}})^{-1} = \mathbf{P} - \mathbf{Q}$
- $\sin \theta = \frac{1}{\|\mathbf{P} - \mathbf{Q}\|_2}$
- $\|\mathbf{P} - \mathbf{Q}\|_2 = \|\mathbf{P}\|_2$

Proof: The first equation can be derived from (4.3), or it can be verified by direct multiplication. The second and third equations follow from the first in conjunction with the results of Theorems 3.1 and 4.1. ■

Corollary 5.6. *For complementary spaces $\mathcal{R}, \mathcal{N} \subset \mathbb{R}^n$, the oblique projector \mathbf{P} onto \mathcal{R} along \mathcal{N} is given by the pseudoinverse of $\mathbf{P}_{\mathcal{N}^\perp} \mathbf{P}_{\mathcal{R}}$ where $\mathbf{P}_{\mathcal{R}}$ and $\mathbf{P}_{\mathcal{N}^\perp}$ are the orthogonal projectors onto \mathcal{R} and \mathcal{N}^\perp , respectively. That is*

$$\mathbf{P} = (\mathbf{P}_{\mathcal{N}^\perp} \mathbf{P}_{\mathcal{R}})^\dagger.$$

Furthermore, if $\bar{\theta}$ is the minimal angle between \mathcal{R} and \mathcal{N}^\perp , then

$$\cos \bar{\theta} = \|\mathbf{P}^\dagger\|_2.$$

Proof: To obtain the first equality, use (3.2) together with $\mathbf{C}^{-1} = \mathbf{V}_1^T \mathbf{U}_1$ to write

$$\mathbf{P}^\dagger = \mathbf{V} \begin{pmatrix} \mathbf{C}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{U}^T = \mathbf{V}_1 \mathbf{C}^{-1} \mathbf{U}_1^T = \mathbf{V}_1 \mathbf{V}_1^T \mathbf{U}_1 \mathbf{U}_1^T = \mathbf{P}_{\mathcal{N}^\perp} \mathbf{P}_{\mathcal{R}}.$$

Now take the pseudoinverse of both sides (see [3] for details concerning pseudoinverses). The second equality is a consequence of the first in conjunction with Theorem 2.1. ■

ACKNOWLEDGMENTS. We thank Steve Campbell for insightful discussions as well as the referee for suggestions that improved the exposition of our paper.

REFERENCES

- [1]. S. N. Afriat, *Orthogonal and oblique projectors and the characteristics of pairs of vector spaces*, Proc. Cambridge Philos. Soc., 53 (1957), pp. 800–816.
- [2]. A. Björck and G. H. Golub, *Numerical methods for computing angles between linear subspaces*, Math. Comp., 27 (1973), pp. 579–594.
- [3]. S. L. Campbell and C. D. Meyer, *Generalized Inverses of Linear Transformations*, Dover Publications (1979 edition by Pitman Pub. Ltd., London), New York, 1991.
- [4]. C. Davis and W. M. Kahan, *The rotation of eigenvectors by a perturbation, III*, SIAM J. Numer. Anal., 7 (1970), pp. 1–46.
- [5]. I. C. Gohberg and M. G. Kreĭn, *Introduction to the Theory of Linear Nonselfadjoint Operators*, American Mathematical Society, Translations of Mathematical Monographs, Vol. 18, Providence, RI, 1969.
- [6]. G. H. Golub and C. F. Van Loan, *Matrix Computations, Second Ed.*, The Johns Hopkins Press, Baltimore, 1989.
- [7]. T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.
- [8]. V. E. Ljance, *Some properties of idempotent operators*, Teor. i Prikl. Mat. L'vov, 1 (1959), pp. 16–22, (Russian).
- [9]. J. M. Schumacher, *A pointwise criterion for controller robustness*, Systems and Control Letters, 18 (1992), pp. 1–8.
- [10]. G. W. Stewart, *Error and perturbation bounds for subspaces associated with certain eigenvalue problems*, SIAM Rev., 15 (1973), pp. 727–764.
- [11]. ———, *On the perturbation of pseudo-inverses, projections, and linear least squares problems*, SIAM Rev., 19 (1977), pp. 634–662.
- [12]. G. W. Stewart and J. Sun, *Matrix Perturbation Theory*, Academic Press, Boston, 1990.
- [13]. G. Strang, *The fundamental theorem of linear algebra*, American Mathematical Monthly, 100 (1993), pp. 848–855.
- [14]. ———, *Linear Algebra and Its Applications, Third Ed.*, Harcourt, Brace, and Jovanovich, San Diego, 1988.

- [15]. P. Wedin, *On angles between subspaces of a finite dimensional inner product space*, in Matrix Pencils, B. Kagstrom and A. Ruhe, eds., Lecture Notes in Mathematics, No. 973, Springer-Verlag, Berlin, Heidelberg, New York, 1982, pp. 263–285.
- [16]. S. Q. Zhu, *Robust complementarity and its application to robust stabilization*, preprint from the Automation and Robotics Research Institute, University of Texas at Arlington, Fort Worth, Texas.

Mathematics Department
North Carolina State University
Raleigh, NC 27695-8205
ipsen@math.ncsu.edu
meyer@math.ncsu.edu

We often hear that mathematics consists mainly of “proving theorems.” Is a writer’s job mainly that of “writing sentences”?

—*Gian-carlo Rota*

In preface to “*The Mathematical Experience*”. Philip J. Davis
and Reuben Hersh. Boston: Birkhäuser, 1981.

NOTES

Edited by: John Duncan

The Four-Vertex Theorem Revisited—Two Variations on the Old Theme

Serge Tabachnikov

The classical four-vertex theorem states that a closed imbedded smooth plane curve has at least four vertices; a vertex is an extremum of curvature. There are many proofs of this theorem—see, e.g. [B-G, O] and the references therein. In some recent works the four-vertex theorem was considered and generalized from the viewpoint of symplectic topology and Sturm theory—see [A1-4, T, G-M-O, O-T]. We present here two results inspired by the four-vertex theorem.

With a smooth plane curve γ another curve Γ , called its caustic (or evolute) is associated: Γ is the envelope of the family of normal lines to γ . Generically the curvature extrema of γ are simple maxima or minima (so that the second derivative of curvature does not vanish at its critical points). Assume that γ is in general position in this sense. Then Γ is a smooth front, that is a singular curve such that at each point the tangent line is well defined—see Fig. 1 (in more technical terms, a front is the projection of a smooth Legendrian curve in the contact manifold of contact elements of the plane to this plane; the general position condition means that the Legendrian curve has only simple tangency with the fibers of the projection—see [A2] for details). Singularities of Γ correspond to vertices of γ ; generically they are cusps shown in the figure. The four-vertex theorem states that the caustic of a closed imbedded curve has at least four cusps.

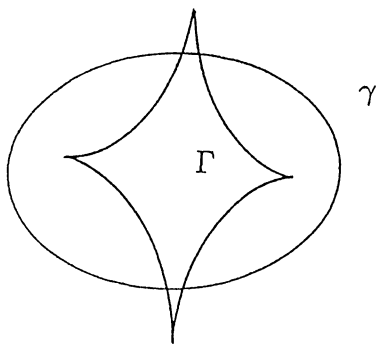


Figure 1

Parametrize the curve γ by length parameter; we write $\gamma(t)$ and use primes to indicate the derivative with respect to t . Then the normal line to γ at point $\gamma(t)$ is

generated by the acceleration vector $\gamma''(t)$. We formulate the four-vertex theorem once again: the envelope of the family of lines through points $\gamma(t)$ in the directions of $\gamma''(t)$ has at least four cusps. This is the statement we generalize in our first theorem.

Consider a smooth plane curve γ . Call its parametrization (not necessarily by length) *definite* if the acceleration vector revolves all the time in the same sense; analytically: $\gamma'''(t) \wedge \gamma''(t) \neq 0$ for all t , where \wedge denotes the determinant of two vectors. Let Γ be the envelope of the family of lines $l(t)$ through points $\gamma(t)$ in the direction of $\gamma''(t)$. As before Γ is a front; the points of γ corresponding to its cusps will be referred to as (generalized) vertices.

Theorem 1. *A generic convex closed smooth curve with a definite parametrization has at least four generalized vertices.*

Proof: To start with we claim that through every point x in the plane at least one (actually two) of the lines $l(t)$ passes. Indeed, consider the function $f(t) = (\gamma(t) - x) \wedge \gamma'(t)$. This function on the circle has a maximum at some point t_0 . Then $0 = f'(t_0) = (\gamma(t_0) - x) \wedge \gamma''(t_0)$, that is the vectors $\gamma(t_0) - x$ and $\gamma''(t_0)$ are collinear. Thus $x \in l(t_0)$. Since the lines $l(t)$ are tangent to Γ we conclude that from each point in the plane there exists a tangent line to Γ .

Next consider the front Γ . It is oriented by the vectors $\gamma''(t)$. Since the parametrization of γ is definite, the tangent direction to Γ revolves in the same sense all the time, and its total turn is 2π . That is, the Gauss map of Γ is one-to-one. If Γ has no cusps then it is a closed convex curve, and there are no tangent lines to Γ from points inside it. This contradicts the previous paragraph.

Alternatively, a somewhat messy computation, which we omit, shows that vertices of γ are critical points of the function $g(t) = \gamma'''(t) \wedge \gamma''(t) / (\gamma''(t) \wedge \gamma'(t))^2$ (if t is the arc length parameter then $g(t)$ is the negative of the curvature). This function on the circle has at least one maximum and one minimum, hence γ has at least two vertices.

Finally we want to show that $g(t)$ has at least two local maxima and two local minima. Suppose not; then Γ has only two cusps. Consider a locally constant function $\phi(x)$ in the complement of Γ whose value at point x equals the number of tangent lines to Γ through x . The value of this function increases by 2 as x crosses Γ from the locally concave to the locally convex side—see Fig. 2. Let x be sufficiently far away from Γ . Since the Gauss map is one-to-one, $\phi(x) = 2$ (indeed there exist exactly two tangent lines to Γ from every point of the circle at infinity; by continuity the same holds for sufficiently distant points x).

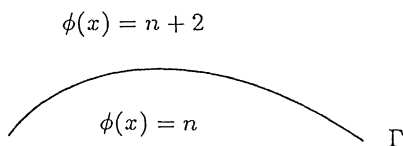


Figure 2

Consider the line through two cusps of Γ (which well may coincide); assume it is horizontal—see Fig. 3. Then the height function restricted to Γ attains either minimum or maximum (or both) not in a cusp. Assume it is maximum; draw the horizontal line l through it. Since Γ lies below this line, $\phi = 2$ above it. Therefore

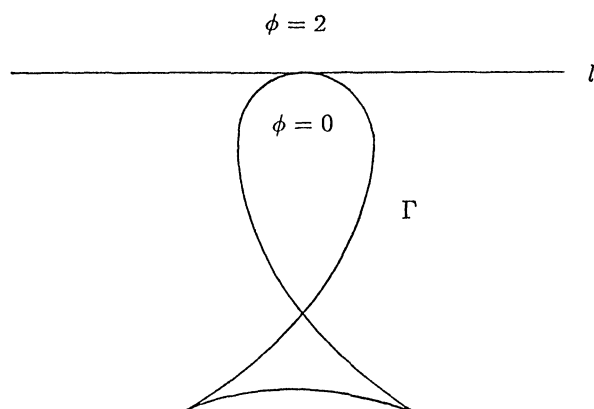


Figure 3

$\phi(x) = 0$ immediately below l , and there are no tangent lines to Γ from x . This contradicts the first paragraph of the proof. Q.E.D.

We pose some questions. First, is the assumption that the parametrization is definite really needed? If it fails, the front Γ will have branches that go to infinity.

Secondly, can one show that the function $g(t)$ has “many” critical points analytically? An interesting example is the affine parametrization characterized by $\gamma''(t) \wedge \gamma'(t) = 1$ for all t (it is not necessarily definite). In this case there exist at least six affine vertices—see [G-M-O] for a modern treatment.

Thirdly, call a generalized diameter of $\gamma(t)$ a chord which is collinear with the acceleration vectors $\gamma''(t)$ at both end points. Said differently, a generalized diameter is a double tangent line of Γ . Does γ always have at least 2 diameters? It is the case for the length parametrization.

Now we proceed to the second theorem. Given a closed plane curve γ we are interested in the following *tripod* configurations: three perpendiculars to γ dropped from one point that make angles of $2\pi/3$ —see Fig. 4.

Theorem 2. *For any smooth convex closed curve there exist at least two tripod configurations.*

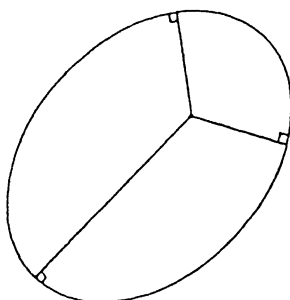


Figure 4

Proof: Let $\gamma(t)$ be a length parametrization, and let $\alpha(t)$ be the angle made by $\gamma'(t)$ with a fixed direction. Because of convexity we may (and will) use α as a parameter on γ . The curvature of γ is $\alpha'(t) = |\gamma''(t)|$.

Choose an origin O inside γ and let $p(t)$ be the signed distance from O to the line $l(t)$, i.e. $p = \gamma \wedge \gamma'' / |\gamma''|$. Let $q = \gamma \wedge \gamma'$; we consider it as a function of α . One has:

$$\frac{dq}{d\alpha} = \frac{dq}{dt} \frac{dt}{d\alpha} = \frac{\gamma \wedge \gamma''}{|\gamma''|} = p.$$

Now consider the function $q(\alpha - 2\pi/3) + q(\alpha) + q(\alpha + 2\pi/3)$. It has a minimum and a maximum on the circle, say at points α_1 and α_2 . Since p is the derivative of q we have:

$$p(\alpha_i - 2\pi/3) + p(\alpha_i) + p(\alpha_i + 2\pi/3) = 0, \quad i = 1, 2.$$

Consider three lines $l(\alpha - 2\pi/3)$, $l(\alpha)$ and $l(\alpha + 2\pi/3)$. They make an equilateral triangle; let a be the length of its side and A its area. Then

$$a(p(\alpha - 2\pi/3) + p(\alpha) + p(\alpha + 2\pi/3)) = 2A$$

—see Fig. 5 (it does not matter whether the origin lies inside or outside the triangle). For α_1, α_2 the left-hand side vanishes and the triangle degenerates to a point. We obtain two tripods. Q.E.D.

The reader may remember a version of the tripod theorem from his school years: there exists a point inside a triangle, whose angles are less than $2\pi/3$, from which all sides are seen at angles $2\pi/3$. This point minimizes the sum of distances to the vertices.

We conclude with another question: can the convexity assumption in the tripod theorem be relaxed? Does it hold for self-intersecting curves?

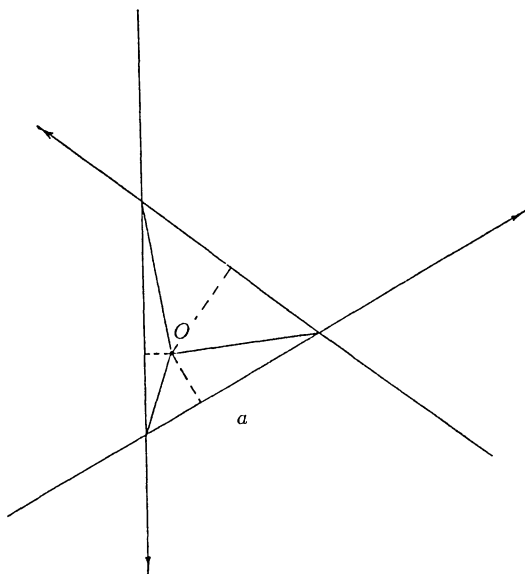


Figure 5

ACKNOWLEDGMENTS. The author is grateful to V. Arnold and V. Ovsienko for stimulating discussions and to the referee for useful suggestions. It is a pleasure to acknowledge the hospitality of the Wolfson College and the I. Newton Institute at the University of Cambridge.

REFERENCES

- [A1] V. Arnold. Ramified Covering $\mathbb{CP}^2 \rightarrow \mathbb{S}^4$, Hyperbolicity and Projective Topology. *Sib. Math. J.*, 29, 1988, No 5, 36–47.
- [A2] V. Arnold. Topological Invariants of Plane Curves and Caustics. *AMS Univ. Lect. Ser.*, 5, Providence, 1994.
- [A3] V. Arnold. On the Number of Flattening Points of Space Curves. Preprint, 1994.
- [A4] V. Arnold. Remarks on the Sextactic and Other Points of Plane Curves. Preprint, 1995.
- [B-G] M. Berger, B. Gostiaux. *Differential geometry: Manifolds, Curves, and Surfaces*. Springer-Verlag, 1988.
- [G-M-O] L. Guieu, E. Mourre, V. Ovsienko. Theorem on Six Vertices of a Plane Curve via Sturm Theory. Preprint, 1994.
- [O] R. Osserman. The Four-or-More Vertex Theorem. *Amer. Math. Monthly*, 92, 1985, 332–337.
- [O-T] V. Ovsienko, S. Tabachnikov. Sturm Theory, Ghys Theorem on Zeroes of the Schwarzian Derivative and Flattening of Legendrian Curves. Preprint, 1995.
- [T] S. Tabachnikov. Around Four Vertices. *Russ. Math. Surv.*, 45, 1990, No 1, 229–230.

Department of Mathematical Sciences
University of Arkansas
Fayetteville, AR 72701
and
Wolfson College and I. Newton Institute
University of Cambridge
16 Mill Lane
Cambridge, CB2 1SB, ENGLAND

Entire Functions Which Vanish at Infinity

R. B. Burckel

In a recent article in this journal David Armitage [1] showed how the pole-shoving technique used in 1885 by Carl Runge, and in most textbook accounts today of Runge's approximation theorem, can be exploited to prove a non-uniqueness theorem for the Radon transform. The idea is to push poles out to infinity through the region S between two confocal parabolas, while keeping the functions small outside S . Since every straight line spends only a compact amount of time in S , the limit entire function g vanishes at infinity along every line, although it is not identically 0. This is quite startling in view of the fact (maximum modulus principle) that the vanishing condition $\lim_{r \rightarrow +\infty} \sup_{|z|=r} |F(z)| = 0$ forces an entire function F to be 0. If one chooses a "transcendental swath" $S := \{x + iy : x \geq 1, ae^x \leq y \leq be^x\}$ ($0 < a < b$), then the non-zero entire function g produced even vanishes at infinity along every unbounded algebraic curve, since every such curve must escape from this S . This clever idea goes back to Harold Bohr [2].

I'd like to show the reader how slight modifications in Armitage's development prove this dramatic result of Bohr. More results of this flavor can be found in Burckel and Saeki [3], which also contains a more extensive bibliography.

First of all, we can take over Armitage's pole-shoving lemma without change:

Lemma 1. Suppose $z_1, z_2 \in \mathbb{C}$, $0 < |z_1 - z_2| < 1$, ϕ_1 is holomorphic in $\mathbb{C} \setminus \{z_1\}$ and $\epsilon > 0$. Then there exists a holomorphic function ϕ_2 in $\mathbb{C} \setminus \{z_2\}$ which satisfies

$$|\phi_1(z) - \phi_2(z)| \leq \epsilon(1 + |z|)^{-2} \quad \text{whenever } |z - z_2| \geq 1.$$

Given $0 < a < b < +\infty$, write

$$S = S_{a,b} := \{(x, y) \in \mathbb{R}^2 : x \geq 1, ae^x \leq y \leq be^x\}$$

and consider any non-zero polynomial $P(X, Y) \in \mathbb{C}[X, Y]$. Then we have the following elementary

Lemma 2. The set $S \cap P^{-1}(0)$ is bounded.

Proof: There are polynomials $Q_0(X), \dots, Q_n(X) \in \mathbb{C}[X]$ such that

$$P(X, Y) = Q_n(X) + Q_{n-1}(X)Y + \dots + Q_0(X)Y^n.$$

We may assume that $Q_0(X)$ is not 0 and write $Q_0(X) = \sum_{j=0}^N c_j X^{N-j}$ with $c_0 \neq 0$. Consider $(x, y) \in S \cap P^{-1}(0)$. Then $x \geq 1$ and $y > 0$, and

$$0 = \frac{P(x, y)}{c_0 x^{N-1} y^n} = \sum_{j=1}^n \frac{Q_j(x)}{c_0 x^{N-1} y^j} + \sum_{j=1}^N \frac{c_j}{c_0 x^{j-1}} + x,$$

whence

$$|x| \leq \sum_{j=1}^n \frac{|Q_j(x)|}{|c_0|(ae^x)^j} + \sum_{j=1}^N \frac{|c_j|}{|c_0|}, \quad (*)$$

since $x \geq 1$ and $y \geq ae^x$. Since $\lim_{x \rightarrow +\infty} x^k e^{-x} = 0$ for every $k \in \mathbb{N}$, the right-hand side of inequality (*) is a bounded function of $x \in [1, +\infty[$. If B is a finite bound, then (*) shows that $|x| \leq B$, and then $|y| \leq be^x \leq be^B$.

The Construction. Special notation for the sequel is

$$z_k := \log(k/3) + 2ik/3 \quad \text{for all integers } k \geq 81,$$

$$g_{81}(z) := \left(\frac{z_{81}}{z - z_{81}} \right)^2, \quad z \in \mathbb{C} \setminus \{z_{81}\},$$

$$S := \{(x, y) \in \mathbb{R}^2 : x \geq 1, e^x/9 \leq y \leq 27e^x\}.$$

The reader can easily check that

$$|z_k - z_{k-1}| < 1 \quad \forall k > 81 \quad (1)$$

and, using only the crude estimate $e < 3$,

$$D(z_k, 2) := \{z \in \mathbb{C} : |z - z_k| \leq 2\} \subset S \quad \forall k \geq 81. \quad (2)$$

Just as in Armitage [1], (1) and Lemma 1 let us inductively construct, for $k > 81$, functions g_k holomorphic in $\mathbb{C} \setminus \{z_k\}$ that satisfy

$$|g_k(z) - g_{k-1}(z)| \leq 2^{-k}(1 + |z|)^{-2} \quad \forall z \in \mathbb{C} \setminus D(z_k, 1). \quad (3)$$

As there, the sequence $\{g_k\}_{k \geq 81}$ converges uniformly on each compact subset of \mathbb{C} to an entire function g which is not constant and which satisfies

$$|g(z)| \leq 8|z_{81}|^2|z|^{-2} \quad \forall z \in \mathbb{C} \setminus \bigcup_{k \geq 81} D(z_k, 1) \setminus D(0, 2|z_{81}|). \quad (4)$$

The Cauchy estimates let us infer from (4) useful bounds on the derivatives of g as well:

$$|g^{(n)}(\xi)| \leq 32n!|z_{81}|^2|\xi|^{-2} \quad \forall \xi \in \mathbb{C} \setminus S \setminus D(0, 2|z_{81}| + 1). \quad (5)$$

Indeed, if $\xi \in \mathbb{C} \setminus S$, then by (2) the whole disk $D(\xi, 1)$ lies in $\mathbb{C} \setminus \bigcup_{k \geq 81} D(z_k, 1)$ and therefore (4) yields

$$|g(z)| \leq 8|z_{81}|^2|z|^{-2} \leq 8|z_{81}|^2(|\xi| - 1)^{-2} \leq 32|z_{81}|^2|\xi|^{-2}$$

for all $z \in D(\xi, 1)$, provided $|\xi| > 2|z_{81}| + 1$ (to insure that $|z| > 2|z_{81}|$ and that $|\xi| - 1 \geq \frac{1}{2}|\xi|$). Now apply the Cauchy estimates in the disk $D(\xi, 1)$.

Conclusion. As Armitage pointed out, the non-zero entire function g' has the interesting property that

$$\int_L |g'| < +\infty \quad \text{and} \quad \int_L g' = 0 \quad \text{for every straight line } L \text{ in } \mathbb{C}.$$

But g also has another remarkable property. Let $\Gamma: [0, +\infty[\rightarrow \mathbb{C}$ be continuous and satisfy $\lim_{t \rightarrow +\infty} |\Gamma(t)| = +\infty$. Suppose in addition that Γ is *algebraic*, that is, it lies in the zero-set of some non-zero polynomial $P(X, Y) \in \mathbb{C}[X, Y]$. For every such Γ and every $n \in \mathbb{N} \cup \{0\}$

$$\lim_{t \rightarrow +\infty} g^{(n)}(\Gamma(t)) = 0.$$

Indeed, as $t \rightarrow +\infty$ the modulus of the point $\Gamma(t)$ in $P^{-1}(0)$ converges to $+\infty$, so by Lemma 1 this point leaves $S \cup D(0, 2|z_{81}| + 1)$ never to return, and inequality (5) becomes valid for $\xi = \Gamma(t)$.

Final Remark. Suppose $f(t, z)$ is integrable in t and holomorphic in z . Various boundedness conditions on f insure that $\int_0^\infty f(t, z) dt$ is holomorphic in z , and conclusions of this type are indispensable in analysis. However, without some supplemental conditions the integral may fail to be holomorphic. The entire function g above was used by Hayman [4] to construct an example of this: $\int_0^\infty e^{-t} g(tz) dt$ is not holomorphic in any neighborhood of 0.

REFERENCES

1. D. H. Armitage, "A non-constant continuous function on the plane whose integral on every line is zero," *Amer. Math. Monthly* 101 (1994), 892–894.
2. H. Bohr, "Über ganze transzendente Funktionen von einem besonderen Typus, Beispiel einer allgemeinen Konstruktionsmethode," *Sitzungsber. Preuss. Akad. Wissen. Berlin, Phys.-Math. Kl.* 26 (1929), 565–571. This is paper E11 in vol. 3 of *Harold Bohr, Collected Mathematical Papers*, Dansk Matematisk Forening (1952), København.
3. R. B. Burckel and S. Saeki, "Entire functions with spiral limits," *Ann. Acad. Scient. Fennicae Ser. A. I. Math.* 9 (1984), 145–151.
4. W. K. Hayman, "On a non-regular parametric integral," *Proc. Royal Soc. Edin.* A83 (1979), 185–188.
5. C. Runge, "Zur Theorie der eindeutigen analytischen Functionen," *Acta Math.* 6 (1885), 229–244.

*Department of Mathematics
Kansas State University
Manhattan, KS 66506*

A Converse to Cauchy's Inequality

D. Zagier

Denote by \mathfrak{M} the set of monotone decreasing functions $f: [0, \infty) \rightarrow [0, 1]$ for which $I(f) = \int_0^\infty f(x) dx$ converges. For $f, g \in \mathfrak{M}$ the scalar product $(f, g) = I(fg)$ converges, and the Cauchy-Schwarz inequality and the inequalities $0 \leq f(x) \leq 1$ imply the estimate

$$(f, g) \leq \min(I(f), I(g), (f, f)^{1/2}(g, g)^{1/2}) \quad (f, g \in \mathfrak{M}). \quad (1)$$

An inequality for (f, g) in terms of the same data but in the other direction, namely

$$(f, g) \geq \frac{(f, f)(g, g)}{\max(I(f), I(g))} \quad (f, g \in \mathfrak{M}), \quad (2)$$

was proved in an earlier article with the same title (up to translation) by a trick involving a quadruple integral [2]. We give here a more general result with a much simpler proof.

Theorem. *Let f and g be monotone decreasing nonnegative functions on $[0, \infty)$. Then*

$$(f, g) \geq \frac{(f, F)(g, G)}{\max(I(F), I(G))} \quad (3)$$

for any integrable (but not necessarily monotone) functions $F, G: [0, \infty) \rightarrow [0, 1]$.

Proof: For all $x \geq 0$ we have

$$\begin{aligned} (f, F) &= I(F)f(x) + \int_0^\infty [f(t) - f(x)]F(t) dt \\ &\leq I(F)f(x) + \int_0^x [f(t) - f(x)] dt, \end{aligned}$$

and hence, since $\int_0^x G(t) dt$ is bounded from above by both x and $I(G)$,

$$\begin{aligned} (f, F) \int_0^x G(t) dt &\leq I(F)xf(x) + I(G) \int_0^x [f(t) - f(x)] dt \\ &\leq \max(I(F), I(G)) \int_0^x f(t) dt. \end{aligned}$$

Now multiply by $-dg(x)$ and integrate by parts from 0 to ∞ . The left-hand side gives $(f, F)(g, G)$, the right-hand side gives $\max(I(F), I(G))(f, g)$, and the inequality remains true because the measure $-dg(x)$ is nonnegative. ■

Remarks. 1. Another proof of (3) can be obtained as follows. It is geometrically clear (and easily proved) that for f monotone decreasing the largest value of (f, F) as F ranges over integrable functions $[0, \infty) \rightarrow [0, 1]$ with a given value of $I(F)$ is attained by taking F to be “as far left as possible,” i.e., equal to 1 for $0 \leq x \leq I(F)$ and to 0 otherwise. Therefore the maximum of $(f, F)(g, G)/A$ as F and G range

over functions $[0, \infty) \rightarrow [0, 1]$ with $\max(I(F), I(G)) \leq A$ is equal to $A^{(-1)} \int_0^A f(x) dx \cdot \int_0^A g(x) dx$, and this is $\leq (f, g)$ because the average of the product of two decreasing functions on an interval is at least equal to the product of their averages.

2. A further generalization of (2) is the inequality

$$W(fg) \geq \frac{W(fF)W(gG)}{\max(W(F), W(G))}$$

valid for any positive linear functional $W(f) = \int_0^\infty f(x)w(x) dx$ ($w(x) > 0$), monotone decreasing functions f and g , and functions $F, G: [0, \infty) \rightarrow [0, 1]$ with $W(F)$ and $W(G)$ finite. To prove it, apply (3) to the functions $f \circ \nu$, $g \circ \nu$, $F \circ \nu$ and $G \circ \nu$, where $\nu(x) = \int_0^x w(x') dx'$. The case $F = f$, $G = g$ is the weighted generalization of (2) proved in [1].

3. As pointed out in [1], both bounds (1) and (2) are best possible in terms of the four parameters $I(f)$, (f, f) , $I(g)$, and (g, g) . The bound (2) cannot be attained for generic values of these parameters but can be approached arbitrarily closely by taking f and g to be step functions with only two non-zero values (i.e. equal to 1 for $x \leq x_0$, to C for $x_0 < x \leq x_1$, and to 0 for $x > x_1$, where $0 < x_0 < x_1$ and $0 < C < 1$). Such functions with given values of $I(f)$ and (f, f) form a one-parameter family (the numbers x_0 , x_1 and C determine each other). If $I(g) \leq I(f)$ and we let f move to the left ($x_0 \rightarrow 0$) and g to the right ($C \rightarrow 0$) in their respective families, then $\int_0^\infty f(x)g(x) dx$ tends to $(f, f)(g, g)/I(f)$.

4. Monotone decreasing functions $f: [0, \infty) \rightarrow [0, 1]$ can be interpreted as the integrals of probability measures ($f(x) = \int_x^\infty d\mu$ where $d\mu$ is a nonnegative measure with integral 1). Hence (2) can be interpreted as a statement about correlations of statistical distributions. One such result, which was the original motivation for the inequality, is an estimate of the possible values of the "Gini coefficient" for a population consisting of two sub-populations, when the size, average income, and Gini coefficients of each of these is given [3]. (The Gini coefficient is a measure of the inequity of distributions of income in a large population which is used widely in mathematical economics.) Since the inequalities (2) and (3) are very general, they should have other applications, perhaps also in pure mathematics.

ACKNOWLEDGMENTS. I would like to thank Joop Kolk for encouragement and helpful conversations, and the students of Analyse C for their interest and patience.

REFERENCES

1. J. V. Pečarić, *A weighted version of Zagier's inequality*, Nieuw Archief voor Wiskunde **12** (1994), 125–127.
2. D. Zagier, *Een ongelijkheid tegengesteld aan die van Cauchy*, Proc. Nederl. Akad. Wet. **80** (1977), 349–351.
3. D. Zagier, *Inequalities for the Gini coefficient of composite populations*, J. Math. Economics **12** (1983), 103–118.

Max-Planck-Institut für Mathematik
Gottfried-Claren-Straße 26
D-53225 Bonn, GERMANY
and
Faculteit Wiskunde
Universiteit Utrecht
Budapestlaan 6
3508 TA Utrecht, NETHERLANDS

UNSOLVED PROBLEMS

Edited by: Richard Guy & Richard Nowakowski

In this department the MONTHLY presents easily stated unsolved problems dealing with notions ordinarily encountered in undergraduate mathematics. Each problem should be accompanied by relevant references (if any are known to the author) and by a brief description of known partial or related results. Typescripts should be sent to Richard Guy, Department of Mathematics & Statistics, The University of Calgary, Alberta, Canada T2N 1N4.

Monthly Unsolved Problems, 1969–1995

Richard K. Guy and Richard J. Nowakowski

References in brackets are to year and page numbers of this MONTHLY, while dates in parentheses refer to publications listed at the end, and other items are labelled (tbp) if they are likely to be published formally, or as written communications (wrc) if publication plans are not presently known. Dates and pages in brackets are also appended to items in the bibliography indicating where the problem originally appeared in the MONTHLY.

Klee [1970, 63] asked for the maximum length of a d -dimensional snake, where by **snake** is meant a simple circuit in the d -cube which has no chords. If we denote this maximum length (number of edges) by $s(d)$, then Abbott & Katchalski (1991) show the $s(d) \geq 77 \times 2^{d-8}$. Their paper contains a very good bibliography. The previous best upper bound was Solov'jeva's (1987) $s(d) \leq 2^{n-1} \left(1 - \frac{1}{n^2 - 5n + 7}\right)$ for $n \geq 7$. Hunter Snevily (1994) improves this to $s(d) \leq 2^{n-1} \left(1 - \frac{1}{20n - 41}\right)$ for $n \geq 12$ and conjectures that $s(d) \leq 3 \cdot 2^{n-3} + 2$ for $n \geq 5$.

Notice that Currie [1993, 790] refers to Keränen's (1992) solution of the problem mentioned by Brown [1971, 886]: there are no sequences on four symbols which contain no two identically equal consecutive segments. Jeffrey Shallit sends a bibliography of 59 items, only 12 of which are among the 25 in §E21 of the second (1994) edition of *Unsolved Problems in Number Theory*.

Recall that an **Ulam sequence** [1973, 919], $\{a_i\} = (u, v)$ is defined by $a_1 = u$, $a_2 = v$ and, for $n > 2$, a_n is the least integer expressible *uniquely* as the sum of two distinct earlier terms. Cassaigne & Finch (tbp) have proved that all Ulam sequences $(4, v)$, $5 \leq v \equiv 1 \pmod{4}$ have precisely three even terms and hence are **regular** in the sense that their differences are ultimately periodic [1993, 946]. They prove that the asymptotic density $\Delta(v)$ of $(4, v) \rightarrow 0$ as $v \rightarrow \infty$, but is misbehaved since

$$\lim_{\substack{v \rightarrow \infty \\ v \equiv 1 \pmod{4}}} \inf (v/2)^\theta \cdot \Delta(v) = 1/4 < 0.27164 < \lim_{\substack{v \rightarrow \infty \\ v \equiv 1 \pmod{4}}} \sup (v/2)^\theta \cdot \Delta(v)$$

where $\theta = 2 - \log_2 3$. Shirriff & Pickover (wrc) have defined a natural multi-dimensional extension of Ulam sequences in which multiplication replaces addition.

Hatada (1994) has republished his [1986, 628] problem, together with seven other problems involving the n -dimensional simplex; for example, if R is the circumradius and r the inradius, is $R \geq nr$ for $n \geq 2$?

In connexion with Dawson's problem [1989, 31] to find a subset of a square that contains disjoint connected sets A and B each containing two opposite corners, Leroy F. Meyers observes that his solution to MONTHLY problem E1515 [1963, 95] implies a solution to the opposite corners problem. He gives two ways to make such a separation:

(a) The separation of the rectangle $[-2, 2] \times [-1, 1]$ which is easily modified to separate the unit square. Let

$$A = \{(0, 0)\} \cup \{(x, y) : 0 < |x| \leq 1 \text{ and } y = \sin(\pi/x)\}$$

and let B be the complement of A in the closed rectangle. As defined here, all four corners belong to B . But it's easy to extend A to contain the segments from $(1, 0)$ to $(1, 2)$ and from $(-1, 0)$ to $(-1, -2)$.

(b) For the square $[0, 1] \times [0, 1]$ let

$$A = \{(0, 0), (1, 1)\} \cup \{(x, y) : 0 < x < 1 \text{ and } 0 < y < 1 \text{ and } y/x \text{ irrational}\}$$

and let B be the complement of A in the closed square.

These partitions are essentially those used by Meyers in his solutions to problems 10328 [1993, 689] and 10341 [1993, 874].

For a survey of recent results on permutation polynomials [1988, 243; 1993, 71] see Mullen (1993).

In [1989, 129] Clark Carroll asked for polynomials with integer roots whose derivatives all have integer roots. For cubics the answer is known and can be found, for example, in Walter (1987) or in Buddenhagen, Ford & May (1992); see also MONTHLY problem E3221, solved in [1989, 841–842]. For quartics with a repeated root there is an infinity of solutions, given essentially by the rational points on the elliptic curve $y^2 = x^3 - 156x + 560$, **57612** in Cremona (1992).

Ralph Buchholz writes that the quintic case $(3, 1, 1)$ [i.e., a triple root and two distinct ones] remains unsolved as also the quartic case $(1, 1, 1, 1)$. He & Jim MacDougall have looked at $y = x^n(x-1)(x-a)$ for rational a and $n \geq 2$ and tried to force (only) the first two derivatives to have rational roots. This led to an elliptic curve as in the degree 4 case and the rank for $2 \leq n \leq 10$ is non-zero. Unfortunately they have not yet proved that the rank is always non-zero, but they conjecture that it is. He & Kelly (1995) have a paper on this topic.

In [1992, 178] Connett asked if a bottle with an inside perfectly reflecting surface could be designed so that a beam of light shone into it was permanently trapped. If the light emanates from one point, or if all the rays are parallel then such light traps were found by R. J. MacG. Dawson, B. E. McDonald, J. Mycielski and L. Pachter. If the light is diffuse (e.g., on the plane, every point of a linear segment shines in all directions, or in 3-dimensional space, every point of a disc shines in all directions) then no such trap is possible. This was shown by J. Mycielski (wrc).

Klaus Leeb writes that in 1970 he formulated a problem equivalent to Parker's permutation problem [1993, 287, and see 1993, 948] and that it was solved, modulo a conjecture that turned out to be Hall's theorem, by a student, Klaus Winkelmann (1979). Kemperman & Ott (1994) note that Hall (1952) and Fuchs (1958) had

already answered the question: for what functions $h: G \rightarrow G$ on an (additive) abelian group G does the equation $\sigma(x) + \tau(x) = h(x)$, $x \in G$ have as its solutions a pair of permutations σ and τ of G and they give explicit constructions in a number of cases, including that when $h(x) = x$ and G is finite. They also determine the finite groups where σ , τ can be chosen to be automorphisms.

Gunnar Blom is unable to claim any of James Currie's \$100 prizes [1993, 790], but believes that the methods of Blom & Thorburn (1982) will serve to find the generating function for the number of non- r -repetitive words of length n , where a non- r -repetitive word (r even) is one which contains no patterns of length r which can be broken into two identical blocks, e.g. 1111, 0101 ($r = 4$) or 101101, 000000, 011011 ($r = 6$).

David Callan (1995) uses Möbius inversion to settle his own problem [1994, 571] about the permanent of a matrix of cotangents by showing that if J is the n by n matrix of ones and C is the matrix with $c_{jk} = \cot\{(2k - 2j + 1)\pi/2n\}$, then

$$\text{per}(xJ + C) = n! \left[x^n + \frac{n}{2}x^{n-2} + \frac{n(n-2)}{2 \cdot 4}x^{n-4}x^{n-4} + \frac{n(n-2)(n-4)}{2 \cdot 4 \cdot 6}x^{n-6} + \dots \right].$$

In [1994, 1007], and in (1993) we discussed the permutation problem that Cayley called 'Mousetrap', unaware of the parallel work of Mundfrom (1994) who corrects the errors in Steen's recurrences and calculates the numbers of permutations of n cards with 2 as the first hit. [Recall that a shuffled deck of n cards, numbered $1, 2, \dots, n$, are counted from the top. If the card number does not agree with the count number, transfer the card to the bottom of the deck. If it does agree, set the card aside and start counting again from 1. The game is won if all cards are set aside, but lost if the count reaches $n + 1$.]

In connexion with our 'coin-weighing problems' article [1995, 164], readers may be interested in a 'problem of the week' from Macalester College. This tradition was started 27 years ago by the late Joe Konhauser and is currently maintained by Stan Wagon.

Problem #784. A Question of Imbalance.

Five coins are identical in appearance except for labels A , B , C , D , and E . Each coin has a weight different from that of each of the others. Given an equal-arm balance, what is the minimum number of uses of the balance required to rank order the coins by weight?

Of course, this is not an unsolved problem, but readers may be amused by it, and more persistent ones will generalize it to the ranking of n coins. Hallard Croft wrote, a quarter of a century ago:

This problem is due to Steinhaus. What is the least number of weighings, $k(n)$, that will always suffice, and what strategy does one adopt in planning them? Steinhaus conjectures that one of these best strategies will also minimize the **expected** number of weighings necessary, where we assume that all original orderings are equally likely.

On the one hand, since there are $n!$ original possible arrangements, and each weighing can only sort these into 2 sets, we have, by obvious information theory, that

$$K(n) = \lceil \log_2(n!) \rceil + 1 \sim \log_2 e \left\{ \left(n + \frac{1}{2} \right) \ln n - n \right\}$$

is an upper bound for $k(n)$.

On the other hand, Ford & Johnson (1959) describe a very nice strategy, which needs only $\kappa(n)$ weighings, where

$$\kappa(n) = \sum_{k=1}^n \lceil \log_{2\frac{3}{4}} k \rceil = n \lceil \log_{2\frac{3}{4}} n \rceil - \lfloor 2^{\lceil \log_2 6n \rceil} / 3 \rfloor + \lfloor \tfrac{1}{2} \log_2 6n \rfloor.$$

We now have the opposing extreme conjectures:

(A) We can always find a ‘completely economic’ strategy, i.e., $k(n) = K(n)$;

(B) We cannot improve the Ford & Johnson scheme, i.e., $k(n) = \kappa(n)$.

The first case for which (A), (B) diverge is $n = 12$. Remarkably, they re-converge, temporarily, for $n = 20, 21$. Asymptotically the difference is clearly narrow, being sometimes as little as $0.028 \dots n$, sometimes more than $0.110 \dots n$.

Recent [around 1970!—Ed.] computation by M. S. Patterson at Cambridge strongly suggested that $k(12) = 30$, not 29, thus disproving (A). This has been confirmed by Wells. He further believes that $k(13) = 33$, which would disprove (B).

There seems to be some connection between the configuration at any stage and the factors of the number of configurations still available. ‘Information theory’ rules out some strategies: for example, for $n = 12$, we cannot use twice, during the procedure, the technique for arranging 5 balls in order, itself 15/16 efficient, for we have ‘lost’ too much information.

Critical cases occur where 2^m is very close to, and larger than, $n!$, e.g. 5 is quite critical (120 ‘just less than’ 128). We might define ‘criticalness’ as $\{\log_2 n!\}$, where $\{\}$ denotes fractional part. The distribution of this is an interesting analytical problem.

Knuth (1973) gives another Steinhaus reference (1959) which suggests that Stanisław Trybuła & Czen Ping may have anticipated Ford & Johnson.

Halbeisen & Hungerbühler (1995) discuss the general counterfeit coin problem and ask what is the best sequential or nonsequential strategy when more than one coin is counterfeit and of possibly differing weights? In particular, do sequential solutions, where later parts of the strategy depend on earlier weighings, always need less weighings than nonsequential ones?

SUPPLEMENT

As this column has only been appearing in alternate issues we have less to say than usual, and some readers may have felt starved of unsolved problems, so here are a few which might not have appeared had they had to sustain a whole article to themselves.

Bernardo Recamán constructs the sequence $a_1 = 1$, and, for $n \geq 1$, $a_{n+1} = a_n/n$ or $a_n \times n$ according as n divides a_n or does not, so that $a_2 = 1$, $a_3 = 2$, $a_4 = 6$, $a_5 = 24$, $a_6 = 120$, $a_7 = 20$, $a_8 = 140, \dots$. He asks for an estimate of a_n . Clearly $(n-1)!$ is an upper bound, and the product of the primes between $n/2$ and n , for which a good estimate is $e^{n/2}$, is a lower bound. We can improve the latter by noting that if $k < \sqrt{n}$ and p is a prime in the interval $\frac{n}{k+1} < p \leq \frac{n}{k}$, then $p^\alpha \parallel a_{n+1}$ where α, k are of the same parity. In particular $p \mid a_{n+1}$ for $\frac{n}{2} < p \leq \frac{n}{1}$, for $\frac{n}{4} < p \leq \frac{n}{3}$, for $\frac{n}{6} < p \leq \frac{n}{5}, \dots$, and by a theorem of Mertens, their product is

$\sim e^{n \ln 2} = 2^n$. This may (sometimes) be quite a good estimate: for example, $a_{212} \approx 121.7 \times 2^{211}$, although these numbers have 66 decimal digits.

John P. Robertson (1995) relates Martin LaBar's (1984) problem of finding a 3×3 magic square all of whose entries are squares to the problem of finding squares in arithmetic progression, to finding three rational right triangles with the same area and with the squares of their hypotenuses in arithmetic progression, to the congruent number problem [1980, 43], and hence to the family of elliptic curves $y^2 = x^3 - n^2x$. Indeed, if three rational points on such a curve can be found, which are the doubles of other rational points and whose x -coordinates are in arithmetic progression, then the problem is solved. Andrew Bremner gives the following specimen whose diagonals fail to give the magic sum.

$$\begin{array}{ccc} 15^2 & 20^2 & 60^2 \\ 36^2 & 48^2 & 25^2 \\ 52^2 & 39^2 & 0^2 \end{array}$$

John Robertson's father, J. S. Robertson, appears twice in last February's Monthly, on p. 167, where his initials are misprinted as J. A., and on p. 173, where they appear as J. B.

Judah Rominek defines the function $t(n)$ as the number of ways of factoring n , where permutations are not counted as different. For example, $12 = 6 \cdot 2 = 4 \cdot 3 = 3 \cdot 2 \cdot 2$ so that $t(12) = 4$. What can be said about $t(n)$? Can it even be proved that $t(n) \leq n$? Perhaps there is something in the paper of Yang & Wen (1994), not yet consulted.

REFERENCES

- Ralph H. Buchholz and Susan M. Kelly, On rational-derived quartics, *Bull. Austral. Math. Soc.* **51**(1995) 121–132. [1989, 129]
- Jim Buddenhagen, Charles Ford and Mike May, Nice cubic polynomials, Pythagorean triples and the law of cosines, *Math. Mag.* **65**(1992) 244–249. [1989, 129]
- David Callan, On evaluating permanents and a matrix of cotangents, *Linear and Multilinear Algebra*, **38**(1995) 193–205. [1994, 571]
- Julien Cassaigne and Steven Finch, A class of 1-additive sequences and quadratic recurrences, *Experimental Math.*, (submitted). [1973, 919]
- John E. Cremona, *Algorithms for Modular Elliptic Curves*, Cambridge University Press, 1992.
- Lester R. Ford and Selmer M. Johnson, A tournament problem, this MONTHLY, **66**(1959) 387–389; *MR* **21** #1942. [1995, 164]
- L. Fuchs, Ein kombinatorisches Problem bezüglich abelscher Gruppen, *Math. Nachr.*, **18**(1958) 292–297. [1993, 287]
- Richard K. Guy and Richard J. Nowakowski, Mousetrap, in *Combinatorics, Paul Erdős is Eighty*, Keszthely, 1993, Bolyai Society Math. Studies, 1993, 193–205. [1994, 1007]
- Lorenz Halbeisen and Norbert Hungerbühler, The general counterfeit coin problem, *Discrete Math.*, **147**(1995). [1995, 164]
- Marshall Hall, A combinatorial problem on abelian groups, *Proc. Amer. Math. Soc.*, **3**(1952) 584–587. [1993, 287]
- Kazuyuki Hatada, Problems on the n dimensional simplex, in John M. Rassias (editor) *Geometry, Analysis and Mechanics*, World Sci. Pub. Co., 1994, pp. 109–112. [1986, 628]
- J. H. B. Kemperman and Teunis J. Ott, Complementary permutations for abelian groups, *Aequationes Math.*, **48**(1994) 262–282. [1993, 287]
- V. Keränen, Abelian squares are avoidable on 4 letters, in W. Kuich (ed.) *Automata, Languages and Programming*, Springer Lect. Notes Comput. Sci., **623**(1992) 41–52. [1971, 886]
- Donald Ervin Knuth, *The Art of Computer Programming*, Vol. 3 Sorting and Searching, Addison-Wesley, 1973, esp. §5.3. [1995, 164]

- Martin LaBar, Problem 270, *Coll. Math. J.*, **15**(1984) 69. [1995, this issue]
- Gary L. Mullen, Permutation polynomials over finite fields, in *Finite Fields, Coding Theory and Advances in Communications and Computing*, (Las Vegas NV 1991), *Dekker Lect. Notes Pure Appl. Math.*, **141**(1993) 131–151; *MR 94d*:11097. [1988, 243; 1993, 71]
- Daniel J. Mundfrom, A problem in permutations: the game of ‘Mousetrap’, *Europ. J. Combin.*, **15**(1994) 555–560. [1994, 1007]
- John P. Robertson, Magic squares of squares, *Math. Mag.*, **68**(1995) (to appear). [1995, this issue]
- Ken Shiriff and Clifford Pickover, 1-multiplicative Ulam sequences, (in preparation). [1973, 919]
- Hunter S. Snevily, The snake-in-the-box problem: A new upper bound, *Discrete Math.*, **133**(1994) 307–314. [1970, 63]
- F. I. Solov’jeva, An upper bound for the length of a cycle in an n -dimensional cube, *Diskret. Analiz.*, **45**(1987). [1970, 63]
- Hugo Dynoisy Steinhaus, *Mathematical Snapshots*, Oxford Univ. Press, New York 1950, pp. 37–40. [1995, 164]
- Hugo Steinhaus, *Calcutta Math. Soc. Golden Jubilee Commemoration*, **2**(1959) 323–327. [1995, 164]
- Johann Walter, Über ganze rationale Funktionen dritten Grades mit ganzzahligen Koeffizienten, bei denen Nullstellen und Extrema zugleich ganzzahlig sind, *Praxis Math.*, **29**(1987) 489–492. [1989, 129]
- Mark B. Wells, *Proc. Information Processing Congress 65*, **2**(1965) 497–498. [1965, 164]
- Mark B. Wells, *Elements of Combinatorial Computing*, Pergamon, Oxford-New York-Toronto, 1971. [1965, 164]
- Yang Yao-Chi and Wen Ren-Kai, A conjecture and an upper bound on the number of multiplicative partitions of natural numbers, *Math. Appl.* **7**(1994) 390–397. [1995, this issue]

Department of Mathematics
The University of Calgary
Calgary, Alberta
Canada T2N 1N4

Department of Mathematics
Dalhousie University
Halifax, Nova Scotia
Canada B3H 3J5

The biologist can push it back to the original protist, and the chemist can push it back to the crystal, but none of them touch the real question of why or how the thing began at all. The astronomer goes back untold million of years and ends in gas and emptiness, and then the mathematician sweeps the whole cosmos into unreality and leaves one with mind as the only thing of which we have any immediate apprehension. *Cogito ergo sum, ergo omnia esse videntur*. All this bother, and we are no further than Descartes. Have you noticed that the astronomers and mathematicians are much the most cheerful people of the lot? I suppose that perpetually contemplating things on so vast a scale makes them feel either that it doesn’t matter a hoot anyway, or that anything so large and elaborate must have some sense in it somewhere.

—*Dorothy L. Sayers*

With Robert Eustace. *The Documents in the Case*. New York: Harper and Row, 1930, p. 54.

PROBLEMS AND SOLUTIONS

Edited by:

Richard T. Bumby, Fred Kochman and Douglas B. West

Proposed problems should be sent to the MONTHLY PROBLEMS address given on the inside front cover. Please include solutions and relevant references. Three copies of all items needed to evaluate the problem should be sent.

Solutions of published problems should arrive at the MONTHLY PROBLEMS address given on the inside front cover before May 31, 1996. If possible, solutions should be typed with double spacing. Two copies suffice. Several solutions may be mailed together, but they should be on separate sheets of paper. The problem number and the solver's name and mailing address should appear on each solution. A mailing label should be included if an acknowledgment is desired.

The published solution is likely to be based on a solution that is complete and correct. Additional information, such as references to other appearances of the problem or its solution, is also welcome.

An asterisk () after the number of a problem, or part of a problem, indicates that no solution is currently available.*

PROBLEMS

10487. *Proposed by William P. Wardlaw, United States Naval Academy, Annapolis, MD.*

Let R be a commutative ring with 1 and let A be an n by n matrix over R . If $\mathbf{x} = \langle x_1 \ x_2 \ \dots \ x_n \rangle$ is a vector with entries in R , let (\mathbf{x}) denote the ideal generated by the entries of \mathbf{x} . Show that $(\mathbf{x}A) = (\mathbf{x})$ for all n -tuples \mathbf{x} over R if and only if A is invertible over R .

10488. *Proposed by Murray S. Klamkin, University of Alberta, Edmonton, Alberta, Canada.*

Determine the extreme values of the sum of the lengths of three concurrent and mutually orthogonal chords of a given sphere of radius R if the point of concurrency is at a distance d from the center.

10489. *Proposed by Frank Schmidt, Arlington, VA.*

Let $f(n)$ be the number of isomorphism classes of connected graphs on n vertices whose automorphism group contains a Sylow 2 subgroup of the symmetric group S_n . For example, $f(3) = 2 = f(4)$. Show that $f(n)$ is an even number for $n \geq 3$.

10490. *Proposed by Seung-Jin Bang, Ajou University, Suwon, Korea.*

Show that

$$\sum_{k=1}^n \frac{(-1)^{k-1}}{k} \binom{n}{k} \sum_{j=1}^k \frac{1}{j} \left(\frac{1}{1} + \frac{1}{2} + \dots + \frac{1}{j} \right) = \sum_{k=1}^n \frac{1}{k^3}$$

for all positive integers n .

10491. *Proposed by Jean-Pierre Grivaux, Lycée Chaptal, Paris, France.*

Let B be an open ball containing the origin in the Euclidean space \mathbb{R}^n , and let V denote its volume. B is cut into 2^n parts by the coordinate hyperplanes

$$\Pi_i = \{(x_1, \dots, x_n) : x_i = 0\}$$

for $i = 1, \dots, n$. Prove that at least 2^{n-1} of these parts have volume at most $V/2^{n-1}$.

10492. *Proposed by William Duke, Mathematical Sciences Research Institute, Berkeley, CA.*

Let n be a positive integer. Show that the only integral polynomials of degree less than n that are real and nonnegative at all n -th roots of unity and have constant term 1 are of the form

$$1 + x^d + x^{2d} + \dots + x^{n-d}$$

with $d|n$, or

$$1 - x^d + x^{2d} - \dots - x^{n-d}$$

with $2d|n$.

10493. *Proposed by Richard P. Stanley, Massachusetts Institute of Technology, Cambridge, MA, and Christophe Reutenauer, Université du Québec à Montréal, Montréal, Canada.*

Fix a positive integer k . Let $f_k(m, n)$ be the number of m -tuples $a = (a_0, a_1, \dots, a_{m-1})$ of integers satisfying: (a) $0 \leq a_i \leq n-1$ for all i , and (b) any k circularly consecutive entries of a (i.e., $a_i, a_{i+1}, \dots, a_{i+k-1}$, where the subscripts are taken modulo m so that they lie between 0 and $m-1$) are all distinct. Show that the generating function

$$F_k(x, n) = \sum_{m \geq 1} f_k(m, n) x^m$$

is a quotient of two polynomials in x and n .

NOTES

(10493) Note that it does not suffice to show that $F_k(x, n)$ is a rational function of x for each fixed n . Consider the case $k = 2$: $f_2(m, n)$ can be shown to be $(n-1)^m + (-1)^m(n-1)$ for $n \geq 1$. From this, it follows that $F_2(x, n) = n(n-1)x^2 / ((1+x)(1-(n-1)x))$.

SOLUTIONS

A Random Replacement Process

10201 [1992, 163]. *Proposed by Gunnar Blom, University of Lund and Lund Institute of Technology, Lund, Sweden.*

An urn contains one white ball and one black ball. Draw a ball at random. With probability $1/2$ return it to the urn; otherwise (again with probability $1/2$) put a ball of the opposite color in the urn. Perform n such drawings in succession. Find the mean and variance of the number X_n of white balls appearing in the n drawings. Find the limiting distribution of $n^{-1/2}(X_n - \mathbf{E}(X_n))$.

Solution I, giving mean and variance, by Peter Griffin, California State University, Sacramento, CA. The mean and variance of X_n are $n/2$ and $n/4 - 1/2 + 1/2^n$ respectively. The limiting distribution of $n^{-1/2}(X_n - \mathbf{E}(X_n))$ is normal, with mean 0 and variance $1/4$.

To prove this, define three possible states for the urn immediately before the n^{th} drawing, namely $S_0 = \{B, B\}$, $S_1 = \{B, W\}$, and $S_2 = \{W, W\}$. The rules for drawing and replacement make the sequence of successive states into a Markov chain with transition matrix

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

By induction the n step transition matrix (for $n > 0$) is

$$P^n = \begin{pmatrix} \frac{1}{4} + \frac{1}{2^{n+1}} & \frac{1}{2} & \frac{1}{4} - \frac{1}{2^{n+1}} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} - \frac{1}{2^{n+1}} & \frac{1}{2} & \frac{1}{4} + \frac{1}{2^{n+1}} \end{pmatrix}$$

The initial distribution of states is $(0, 1, 0)$; hence the distribution at all later times is $(1/4, 1/2, 1/4)$

Define $Y_n = 1$ or 0 according to whether the n^{th} ball drawn is white or black. Then $\mathbf{E}(X_n) = \sum_{i=1}^n \mathbf{E}(Y_i)$, and $\mathbf{E}(Y_i) = 1/2$, using the expressions for the distribution of states. Thus $\mathbf{E}(X_n) = n/2$ as might have been anticipated by symmetry.

To compute $\text{Var}(X_n) = \sum_{i=1}^n \text{Var}(Y_i) + 2 \sum_{1 \leq i < j \leq n} \text{Cov}(Y_i, Y_j)$, observe first that $\text{Var}(Y_i) = 1/4$. Also $\text{Cov}(Y_i, Y_j) = \mathbf{E}(Y_i Y_j) - (1/2)^2 = \text{Pr}(Y_i Y_j = 1) - 1/4$. Let S be the state just before the i^{th} drawing, S' the state just before the $(i+1)^{\text{th}}$ drawing, and S'' the state just before the j^{th} drawing. Then

$$\begin{aligned} \text{Pr}(Y_i Y_j = 1) &= \sum_{k,l,m=0}^2 \text{Pr}(Y_i Y_j = 1, S = k, S' = l, S'' = m) \\ &= \sum_{k,l,m} \text{Pr}(Y_j = 1 \mid S'' = m) \cdot \text{Pr}(S'' = m \mid S' = l) \\ &\quad \cdot \text{Pr}(S' = l, Y_i = 1 \mid S = k) \cdot \text{Pr}(S = k) \end{aligned}$$

where we have used an implicit independence of past events, which leads to the strong Markov property of this process, to write $\text{Pr}(Y_j = 1 \mid S'' = m, S' = l, S = k, Y_i = 1)$ as $\text{Pr}(Y_j = 1 \mid S'' = m)$ and $\text{Pr}(S'' = m \mid S' = l, S = k, Y_i = 1)$ as $\text{Pr}(S'' = m \mid S' = l)$.

The terms $Pr(Y_j = 1 \mid S'' = m)$ and $Pr(S' = l, Y_i = 1 \mid S = k)$ are immediately computed from the drawing and replacement rules; the term $Pr(S'' = m \mid S' = l)$ is the (l, m) entry of P^n for $n = j - i - 1$ if $j - i > 1$, otherwise the Kronecker δ_{lm} .

The term $Pr(S = k)$ is the k^{th} entry of the state vector, so there are two cases depending on whether $i = 1$ or $i > 1$. For $i = 1$ we find $Pr(Y_1 Y_j = 1) = 1/4 - 1/2^{j+1}$; for $i > 1$ we find $Pr(Y_i Y_j = 1) = 1/4$. (In particular, for $i > 1$, Y_i and Y_j are independent.) Thus finally

$$\text{Var}(X_n) = \frac{n}{4} - 2 \sum_{j=2}^n \frac{1}{2^{j+1}} = \frac{n}{4} - \frac{1}{2} + \frac{1}{2^n}.$$

Solution II, establishing asymptotic distribution, by Wolfgang J. Bühler, Johannes Gutenberg-Universität, Mainz, Germany. The asymptotics of the process are most easily seen by observing that after each turn a white ball is returned to the urn with probability $1/2$, independently of anything else that has happened. Therefore, the number Z_n of white balls returned to the urn after the first n drawings is the sum of a fair coin tossing process, to which the central limit theorem applies. Since X_n and Z_n can differ by at most 1, $(X_n - \mathbf{E}(X_n)) / \sqrt{n}$ and $(Z_n - \mathbf{E}(Z_n)) / \sqrt{n}$ have the same asymptotic distribution, by lemma 2 of [2, section VIII.2]. Note that this approach shows that the initial distribution does not affect the asymptotic behavior of the process.

Editorial comment. The selections above are extracted from more complete solutions that were submitted. For example, Peter Griffin's method led to the use of the mixing central limit theorem (Theorem 27.5 of [1]) to obtain the asymptotic distribution. In particular, it is not necessary to invent an independent process in order to apply a weaker theorem. The applicability here depends on the fact that P^2 has all its entries positive, and an argument very similar to that of example 27.5 of [1]. To allow for the fact that the initial distribution is not the stationary distribution of the chain, follow the argument of Theorem 19.1.1 of [7].

Ellen Hertz analyzed the process by distinguishing the positions of the two balls, with the replacement always being put in the vacant position. She introduced the random variable T , the earliest time at which both positions have been touched. Conditioned on T , X_n becomes, essentially, a sum of independent random variables. This simplifies the rest of the analysis.

Generating functions provide an efficient organization of the details of Solution I. The transition matrices lead to linear equations relating the generating functions for the different states. Robin Chapman included solutions of these equations. Surya Narayana noted that a general method along these lines has been developed in [3]. Other references giving details and examples of the use of generating functions to establish asymptotic normality can be found in [4], [5], and [6].

REFERENCES

1. P. Billingsley, *Probability and Measure*, Wiley, 1979
2. W. Feller, *An Introduction to Probability Theory and its Applications*, Vol. II (second ed.), Wiley, 1971
3. D. M. Lucantoni, "New results on the single server queue with batch Markovian arrival process", *Stochastic Models* 7 (1991), 1–46
4. E. A. Bender, "Central and local limit theorems applied to asymptotic enumeration", *J. Combin. Theory Ser. A* 15 (1973), 91–111
5. E. A. Bender & L. B. Richmond, "Central and local limit theorems applied to asymptotic enumeration, II", *J. Combin. Theory Ser. A* 34 (1983), 255–265
6. E. A. Bender, L. B. Richmond & S. G. Williamson, "Central and local limit theorems applied to asymptotic enumeration, III: Matrix recursions", *J. Combin. Theory Ser. A* 35 (1983), 263–278
7. I. Ibragimov & Yu. V. Linnik, *Independent and Stationary Sequences of Random Variables*, Wolters-Noordhoff, 1971

Solved also by R. J. Chapman (U. K.), V. Hernández (Spain), E. Hertz, J. H. Lindsey II, S. Narayana, M. Stamp, J. Vogel, and the proposer.

A Solitaire Army on the Line

10287 [1993, 185]. *Proposed by Dr. A. Keith Austin, The University of Sheffield, Sheffield, England.*

We have a doubly-infinite (i.e., indexed by \mathbb{Z}) row of squares and we start with counters in those squares to the left of some point (e.g., those with negative index). For a fixed positive integer k , the allowable moves consist of selecting k consecutive squares, discarding one of the counters in those squares, and rearranging the remaining counters within the k selected squares (with at most one counter in a square). Prove or disprove that there is an integer $N = N(k)$ such that no sequence of moves will allow a counter to be placed N squares into the region which originally contained no counters.

Solution by Ilias Kastanas, California State University, Los Angeles, CA. There is such an integer N . For $k = 1, 2$, we observe that we can never move a counter into non-negative territory. For $k > 2$, let $\lambda > 1$ be a real number satisfying the equation

$$1 + \lambda + \lambda^2 + \dots + \lambda^{\lfloor (k-1)/2 \rfloor} = \lambda^{\lceil (k+1)/2 \rceil} + \dots + \lambda^{k-1}.$$

The defining equation has one more term on the left than on the right. If we delete the same number of terms from each side, what remains on the left exceeds what remains on the right. In particular, the sum of l members of $\{1, \lambda, \dots, \lambda^{k-1}\}$ is at least $\sum_{i=0}^{l-1} \lambda^i$, the sum of $l - 1$ members is at most $\sum_{i=k-l}^{k-1} \lambda^i$, and the former is at least the latter. Hence no move can increase $\sum_{m \in S} \lambda^m$, where S is the set of positions containing counters. Originally, $\sum_{m \in S} \lambda^m = \sum_{m=-\infty}^{-1} \lambda^m = 1/(\lambda - 1)$. If N is large enough so that $\lambda^N > 1/(\lambda - 1)$, then no counter can reach position N .

Solved also by R. J. Chapman (U. K.), B. Doran, W. Goddard & D. J. Kleitman, R. Holzinger, O. P. Lossers (The Netherlands), G. Myerson (Australia), B. Peterson, B. Ravikumar, A. Riese, F. Schmidt, A. N. 't Woord (The Netherlands), GCHQ Problem Solving Group (U. K.), and the National Security Agency Problems Group.

A Rediscovery of Pólya

10290 [1993, 290]. *Proposed by David Allison, University of Cape Town, Rondebosch, South Africa.*

Let $c \in \mathbb{N}$. Consider the expression $S_c(n) = \sum_{r=1}^n r^c$.

- (a) Show that $S_c(n)/S_1^2(n)$ is a polynomial in $S_1(n)$ when c is odd and $c > 1$.
- (b) Show that $S_c(n)/S_2(n)$ is a polynomial in $S_1(n)$ when c is even.

Editorial comment. As the Con Amore Problem Group observed, the statement appears as 3.8 and 3.9 of George Pólya's *Mathematical Discovery, Vol. 1* (Wiley, 1962). Pólya gives, as 3.7 and 3.8, formulas which give the results immediately by induction. The elementary solutions provided by readers used the same method as Pólya. Other solutions were longer, more complicated, or used more advanced results. In particular, these sums may be expressed in terms of values of Bernoulli polynomials, allowing the result to be extracted from many references. The papers of Edwards, cited below, note that work on instances of these identities has been traced back to ancient times, although the idea of writing the S_c in terms of S_1 is credited to Johannes Faulhaber (1580–1635). Authors mentioned below describe their encounter with a copy of his *Academia Algebrae*, published in 1631, in which these formulas are developed. A biography of Faulhaber by Ivo Scheider was published by Birkhäuser in 1993.

Many solvers cited other sources that, if not the exact solution of the problem, provide formulas that gave the result quickly. A sample is given below (as provided by readers in parentheses).

REFERENCES

- O. D. Anderson, "Summing powers of integers", *Mathematical Spectrum* 23 (1990-91), 116–121 (S.-J. Bang).
 B. C. Berndt, *Ramanujan's Notebooks I* (Springer, 1985), 157–158, (H.-J. Seiffert).
 P. Bachmann, *Niedere Zahlentheorie* (Leipzig 1902, 1910), Vol. 2, p. 293, (reprinted by Chelsea, 1968), (M. Vowe).
 S. Bernard and J. M. Child, *Higher Algebra* (Macmillan 1936), VIII.8 (A. Caicedo Núñez).
 A. W. F. Edwards, "Sums of powers of integers", *Mathematical Gazette* 66(1982), 22–28 (F. Flanigan, H. Krishnapriyan).
 A. W. F. Edwards, "A quick route to sums of powers", this MONTHLY 93 (1986), 451–455, (F. Flanigan, A. Pedersen).
 D. E. Knuth, "Johann Faulhaber and sums of powers", *Math. Comp.* 61 (1993), 277–294, (I. Nemes).

Solved also by S.-J. Bang (Korea), R. Barbara (Lebanon), V. Božin (student, Yugoslavia), A. E. Caicedo Núñez (student, Colombia), R. J. Chapman (U. K.), F. J. Flanigan, J. Fukuta (Japan), R. Holzsgager, F. T. Howard, L. N. Howard, H. Kappus (Switzerland), I. Kastanas, B. G. Klein, H. K. Krishnapriyan, S. Liu, O. P. Lossers (The Netherlands), A. D. Melas (Greece), I. Nemes (Austria), A. Pedersen (Denmark), F. Schmidt, H.-J. Seiffert (Germany), R. Stong, M. Vowe (Switzerland), A. N. 't Woord (The Netherlands), Anchorage Math Solutions Group, Con Amore Problem Group (Denmark), and the proposer.

A Large Intersection of Large Sets

10373 [1994, 274]. *Proposed by M. J. Pelling, Balliol College, Oxford, England.*

Let $E_n \subseteq I = [0, 1]$ be a sequence of measurable sets in the unit interval with measures $mE_n \geq \delta > 0$ bounded away from zero. Prove that there is a subsequence $\langle E_{n_i} \rangle$ whose intersection has the cardinality of the continuum.

Solution by Kenneth Schilling, University of Michigan, Flint, MI. By passing to a subsequence we may assume that the sequence $\langle \chi_{E_n} \rangle$ of characteristic functions converges in the weak* topology for, say, L^2 to the measurable function $f: [0, 1] \rightarrow \mathbb{R}$. In particular, then

$$\lim_{n \rightarrow \infty} m(E_n \cap F) = \int_F f$$

for all measurable sets $F \subset [0, 1]$.

Taking $F = [0, 1]$, we see that $\int_0^1 f \geq \delta > 0$; thus there exists $\delta' > 0$ and $G \subset [0, 1]$ such that $m(G) > 0$ and $f > \delta'$ on G . Thus

$$\lim_{n \rightarrow \infty} m(E_n \cap F) \geq \delta' \cdot m(F) \quad (*)$$

for all measurable sets $F \subset G$.

We now construct, by recursion, a family $\langle G_s \rangle$ of closed subsets of G of positive measure, indexed by the elements s of the set Sq of finite binary sequences, and a sequence N_i , such that the follow properties hold for all $s, s' \in Sq$: (1) $G_s \cap G_{s'} = \emptyset$ for $s \neq s'$ of the same length, (2) $G_s \supset G_{s'}$ for $s \subset s'$, and (3) $G_s \subset E_{N_i}$ for s of length i .

Choose N_0 so that $G \cap E_{N_0}$ has positive measure, and let $G_\emptyset = G \cap E_{N_0}$, where \emptyset denotes the empty sequence. Now, suppose that G_s has been suitably defined for all s of length k or less, and N_i has been defined for $i \leq k$. Given $s \in Sq$ of length k , temporarily let A_s and B_s be disjoint subsets of G_s of positive measure. Using (*), there exists $N_{k+1} > N_k$ such that both $E_{N_{k+1}} \cap A_s$ and $E_{N_{k+1}} \cap B_s$ have positive measure for all s of length k . Now, for each s of length k , let G_{s0} and G_{s1} be closed subsets of positive measure of $E_{N_{k+1}} \cap A_s$ and $E_{N_{k+1}} \cap B_s$, respectively. This completes the construction.

For every infinite binary sequence σ , the sets $G_{\sigma_1 \dots \sigma_i}$ are nested, nonempty and compact, so the set $K_\sigma = \bigcap_{i=1}^{\infty} G_{\sigma_1 \dots \sigma_i}$ is nonempty. Also, the sets K_σ are disjoint for distinct σ , so

$\bigcup_{\sigma} K_{\sigma}$ has the cardinality of the continuum. Finally, $G_{\sigma_1 \dots \sigma_i} \subset E_{N_i}$ for all σ and i . Thus, $\bigcup_{\sigma} K_{\sigma} \subset \bigcap_{i=1}^{\infty} E_{N_i}$ and the proof is complete.

Editorial comment. Zbigniew Lipecki noted that stronger and more precise results may be found in P. Erdős, H. Kestenman, and C. A. Rogers, “An intersection property of sets with positive measure”, *Colloq. Math.* 11 (1963), 75–80. A still stronger result (implying, for example, that there is a subsequence whose intersection has Hausdorff dimension 1) can be found in P. Erdős and S. J. Taylor, “The Hausdorff measure of the intersection of sets of positive Lebesgue measure”, *Mathematika* 10 (1963), 1–9. The title of this solution is borrowed from an article by Paul Halmos, this MONTHLY 99 (1992), 307–312 (the author credited credited this article with suggesting the problem, and reference to it was given in a note accompanying the statement of the problem) .

Solved also by R. B. Israel (Canada), I. Krzemińska (Poland), Z. Lipecki (Poland), A. N. 't Woord (The Netherlands), and the proposer.

REVIVALS

A few misprints were noted, which will be corrected here.

Comparing Sums of Numbers with Equal Products

6667 [1991, 766; 1994, 914]. *Proposed by George Baloglou and Phil Tracy, State University of New York, College at Oswego.*

If $a_1, a_2, \dots, a_n, b_1, b_2, \dots, b_n$ are positive numbers such that

$$a_1 a_2 \dots a_n = b_1 b_2 \dots b_n \quad \text{and} \\ \sum_{1 \leq i < j \leq n} |a_i - a_j| \leq \sum_{1 \leq i < j \leq n} |b_i - b_j|,$$

(i) prove that

$$\sum_{i=1}^n a_i \leq (n-1) \sum_{i=1}^n b_i, \quad \text{and}$$

(ii) show that the factor $n-1$ cannot be replaced by a smaller one.

Editorial comment. A mistake that could only be made by a computer took the asterisk indicating that Problem 6080 [1976, 205; 1994, 913] had been initially published without a solution and placed it on part (i) of this problem. The proposers had, indeed, submitted solutions to both parts of this problem. The proposers' lengthy solution (of both (i) and (ii)) relies on a study, for $L > 0$ and $n > 2$, of $\sum_{i=1}^n x_i$ on

$$D(L) = \left\{ (x_1, x_2, \dots, x_n) : x_i > 0, x_1 x_2 \dots x_n = 1, \sum_{1 \leq i < j \leq n} |x_i - x_j| = (n-1)L \right\}.$$

Denote the extreme values $\sum_{i=1}^n x_i$ on $D(L)$ by $\max(L)$ and $\min(L)$. The proposers were able to show that $\max(L)/\min(L) < n-1$ and other properties of these quantities essential to solving the problems. However, determination of these extreme values remains open.

INDEX TO VOLUME 102, 1995

THE AMERICAN MATHEMATICAL MONTHLY

TITLE INDEX

- An Abstract Algebra Story, U. Leron and E. Dubinsky, 227
Alan Turing and the Central Limit Theorem, S. L. Zabell, 483
Approximation Algorithms: Good Solutions to Hard Problems, R. Libeskind-Hadas, 57
Areas of Polygons Inscribed in a Circle, D. P. Robbins, 523
The Binary Expansion of $1/p$, A. R. Meijer, 427
Calculus in the Operating Room, P. Toy and S. Wagon, 101
Coin-Weighing Problems, R. K. Guy and R. J. Nowakowski, 164
Communicating Mathematics: Useful Ideas from Computer Science, C. Wells, 397
Continued Fractions, Chebyshev Polynomials, and Chaos, W. Derrick and J. Eidswick, 337
Cosine Products, Fourier Transforms, and Random Sums, K. E. Morrison, 716
Count-Wheels: A Mathematical Problem Arising in Horology, S. H. Weintraub, 310
Curves of Constant Precession, P. D. Scofield, 531
Derivative Polynomials for Tangent and Secant, M. E. Hoffman, 23
Does the Möbius Function Determine Multiplicative Arithmetic?, D. Flath and A. Zulauf, 354
Down With Determinants!, S. Axler, 139
Drums that Sound the Same, S. J. Chapman, 124
Elliptic Curves, J. Stillwell, 831
An Envy-Free Cake Division Protocol, S. J. Brams and A. D. Taylor, 9
The Evolution of Algebra 1800-1870, I. G. Bashmakova and A. N. Rudakov, 266
Experimentation and Conjecture Are Not Enough, D. T. Haimo, 102
Exploring the Brachistochrone Problem, L. Haws and T. Kiser, 328
Exponentiation in Rings, R. H. Redfield, 36
Football Pools --- A Game for Mathematicians, H. Hämäläinen, I. Honkala, S. Litsyn, and P. Östergård, 579
Four Significant Axiomatic Systems and Some of the Issues Associated with Them, S. Mykytiuk and A. Fresh Breezes in the Philosophy of Mathematics, R. Hersh, 589
Gale's Round-Trip Jeep Problem, A. Hausrath, B. Jackson, J. Mitchem, and E. Schmeichel, 299
Geometry and the Foucault Pendulum, J. Oprea, 515
The Great Marble Race: An Assignment Gone Wrong, B. Evans and J. Johnson, 506
Harvard Calculus at Oklahoma State University, K. Johnson, 794
How to Add Fast --- On Average, G. Schay, 725
How to Write a Proof, L. Lamport, 600
How to Teach a Class by the *Modified* Moore Method, D. R. Chalice, 317
A Hyperbolic Plane Coloring and the Simple Group of Order 168, D. Mackenzie, 706
Integer Hexahedra Equivalent to Perfect Boxes, B. E. Peterson and J. H. Jordan, 41
Isometries of the Plane, D. A. Singer, 628
L. C. Larson, 675
The Law of Large Numbers and $\sqrt{2}$, T. M. Liggett and P. Petersen, 31
The Mathematics Portfolio, M. L. Crowley and K. Dunn, 19
Mathematics: Questions and Answers, B. Eckmann (translated by P. Hilton), 685
Missing Real Numbers, C. J. Van Wyk, 260
Monthly Unsolved Problems, 1969-1995, R. K. Guy and R. J. Nowakowski, 921
A Multidimensional Version of Rolle's Theorem, M. Furi and M. Martelli, 243
My Favorite Elliptic Curve: A Tale of Two Types of Triangles, R. K. Guy, 771
Niels Hendrik Abel and Equations of the Fifth Degree, M. I. Rosen, 495
A Nobel Prize in Mathematics, J. E. Morrill, 888
Off to the Races, J. Ondich, 826
On the Geometry of Halley's Method, T. R. Scavo and J. B. Thoo, 417
On Some Applications of Fibonacci Numbers, D. L. Ranum, 640
Order and Chaos on Your Desk, S. Bassein, 409
Pebbling a Chessboard, F. Chung, R. Graham, J. Morrison, and A. Odlyzko, 113
Pick's Formula via the Weierstrass p -Function, R. Diaz and S. Robins, 431
Polygonal Rooms Not Illuminated From Every Point, G. W. Tokarsky, 867
Quadratics Representing Primes, N. Boston and M. L. Greenwood, 595
The Rise, Fall, and Possible Transfiguration

of Triangle Geometry: A Mini-History, P. J. Davis, 204
 Searching for Common Generalizations: The Case of Hyperbolic Functions, K. B. Stolarsky, 609
 Shenitzer, 62
 The Significant-Digit Phenomenon, T. P. Hill, 322
 Some Exact Number Theory Computations via Probability Mechanisms, R. Blecksmith and P. W. Laud, 893
 Some Problems Concerning Recurrence Sequences, G. Myerson and A. J. Van Der Poorten, 698
 A Spigot Algorithm for the Digits of π , S. Rabinowitz and S. Wagon, 195
 Stalking the Wild Ellipse, K. M. Kendig, 782
 A Story of Binomial Coefficients and Primes, J. W. Sander, 802
 Teaching Math More Effectively, Through Computational Proofs, D. Gries and F. B. Schneider, 691
 The Stochastic Group, D. G. Poole, 798
 The Angle Between Complementary Subspaces, I. C. F. Ipsen and C. D. Meyer, 904
 The Fifty-Fifth William Lowell Putnam Mathematical Competition, L. F. Klosinski, G. L. Alexanderson, and The Role of

Transitivity in Devaney's Definition of Chaos, A. Crannell, 788
 Three Open Problems in Functional Equations, P. K. Sahoo, 741
 Three Sing-Sing Problems, G. Blom, L. Holst, and D. Sandell, 880
 Topology and Abstract Algebra as Two Roads of Mathematical Comprehension, Part II, H. Weyl, 646
 Topology and Abstract Algebra as Two Roads of Mathematical Comprehension, Part I, H. Weyl, 453
 Totally Real Origami and Impossible Paper Folding, D. Auckly and J. Cleveland, 215
 Transforming n -gons by Folding the Plane, P. Sabinin and M. G. Stone, 620
 Turán's Graph Theorem, M. Aigner, 808
 Veni, Divisi, Vici, C. C. McGeoch, 449
 Wanted: A Bad Matrix, G. H. Meisters, 546
 What is the Worth of Free Casino Credit?, M. Orkin and R. Kakigi, 3
 Why Did George Green Write His Essay of 1828 on Electricity and Magnetism?, I. Grattan-Guinness, 387
 Yueh-Gin Gung and Dr. Charles Yu Award for Distinguished Service to Anneli Lax, I. Niven, 99

NAME INDEX

Aigner, M., Turán's Graph Theorem, 808
 Alexanderson, G. L. *see* Klosinski
 Auckly, D. and J. Cleveland, Totally Real Origami and Impossible Paper Folding, 215
 Axler, S., Down With Determinants!, 139
 Bashmakova, I. G. and A. N. Rudakov, The Evolution of Algebra 1800-1870, 266
 Bassein, S., Order and Chaos on Your Desk, 409
 Blecksmith, R. and P. W. Laud, Some Exact Number Theory Computations via Probability Mechanisms, 893
 Blom, G., L. Holst, and D. Sandell, Three Sing-Sing Problems, 880
 Boston, N. and M. L. Greenwood, Quadratics Representing Primes, 595
 Brams, S. J. and A. D. Taylor, An Envy-Free Cake Division Protocol, 9
 Chalice, D. R., How to Teach a Class by the *Modified* Moore Method, 317
 Chapman, S. J., Drums that Sound the Same, 124
 Chung, F., R. Graham, J. Morrison, and A. Odlyzko, Pebbling a Chessboard, 113
 Cleveland, J. *see* Auckly

Crannell, A., The Role of Transitivity in Devaney's Definition of Chaos, 788
 Crowley, M. L. and K. Dunn, The Mathematics Portfolio, 19
 Davis, P. J., The Rise, Fall, and Possible Transfiguration of Triangle Geometry: A Mini-History, 204
 Derrick, W. and J. Eidswick, Continued Fractions, Chebychev Polynomials, and Chaos, 337
 Diaz, R. and S. Robins, Pick's Formula via the Weierstrass p -Function, 431
 Dubinsky, E. *see* Leron
 Dunn, K. *see* Crowley
 Eckmann, B., Mathematics: Questions and Answers, 685
 Eidswick, J. *see* Derrick
 Evans, B. and J. Johnson, The Great Marble Race: An Assignment Gone Wrong, 506
 Flath, D. and A. Zulauf, Does the Möbius Function Determine Multiplicative Arithmetic?, 354
 Furi, M. and M. Martelli, A Multidimensional Version of Rolle's Theorem, 243
 Graham, R. *see* Chung

- Grattan-Guinness, I., Why Did George Green Write His Essay of 1828 on Electricity and Magnetism?, 387
- Greenwood, M. L. *see Boston*
- Gries, D. and F. B. Schneider, Teaching Math More Effectively, Through Computational Proofs, 691
- Guy, R. K., My Favorite Elliptic Curve: A Tale of Two Types of Triangles, 771
- Guy, R. K. and R. J. Nowakowski, Coin-Weighing Problems, 164
- Guy, R. K. and R. J. Nowakowski, *Monthly* Unsolved Problems, 1969-1995, 912
- Haimo, D. T., Experimentation and Conjecture Are Not Enough, 102
- Hämäläinen, H., I. Honkala, S. Litsyn, and P. Östergård, Football Pools --- A Game for Mathematicians, 579
- Hausrath, A., B. Jackson, J. Mitchem, and E. Schmeichel, Gale's Round-Trip Jeep Problem, 299
- Haws, L. and T. Kiser, Exploring the Brachistochrone Problem, 328
- Hersh, R., Fresh Breezes in the Philosophy of Mathematics, 589
- Hill, T. P., The Significant-Digit Phenomenon, 322
- Hoffman, M. E., Derivative Polynomials for Tangent and Secant, 23
- Holst, L. *see Blom*
- Honkala, I. *see Hämäläinen*
- Ipsen, I. C. F. and C. D. Meyer, *The Angle Between Complementary Subspaces*, 904
- Jackson, B. *see Hausrath*
- Johnson, K., *Harvard Calculus at Oklahoma State University*, 794
- Johnson, J. *see Evans*
- Jordan, J. H. *see Peterson*
- Kakigi, R. *see Orkin*
- Kendig, K. M., *Stalking the Wild Ellipse*, 782
- Kiser, T. *see Haws*
- Klosinski, L. F., G. L. Alexanderson, and L. C. Larson, *The William Lowell Putnam Mathematical Competition*, 675
- Lamport, L., *How to Write a Proof*, 600
- Larson, L. C. *see Klosinski*
- Laud, P. W. *see Blecksmith*
- Leron, U. and E. Dubinsky, *An Abstract Algebra Story*, 227
- Libeskind-Hadas, R., *Approximation Algorithms: Good Solutions to Hard Problems*, 57
- Liggett, T. M. and P. Petersen, *The Law of Large Numbers and %2, 31*
- Litsyn, S. *see Hämäläinen*
- Mackenzie, D., A Hyperbolic Plane Coloring and the Simple Group of Order 168, 706
- Martelli, M. *see Furi*
- McGeoch, C. C., Veni, Divisi, Vici, 449
- Meijer, A. R., The Binary Expansion of $1/p$, 427
- Meisters, G. H., Wanted: A Bad Matrix, 546
- Meyer, C. D. *see Ipsen*
- Mitchem, J. *see Hausrath*
- Morrill, J. E., A Nobel Prize in Mathematics, 888
- Morrison, K. E., Cosine Products, Fourier Transforms, and Random Sums, 716
- Morrison, J. *see Chung*
- Myerson, G. and A. J. Van Der Poorten, Some Problems Concerning Recurrence Sequences, 698
- Mykytiuk, S. and A. Shenitzer, Four Significant Axiomatic Systems and Some of the Issues Associated with Them, 62
- Niven, I., Yueh-Gin Gung and Dr. Charles Yu Award for Distinguished Service to Anneli Lax, 99
- Nowakowski, R. J. *see Guy*
- Nowakowski, R. J. *see Guy*
- Odlyzko, A. *see Chung*
- Ondich, J., Off to the Races, 826
- Oprea, J., Geometry and the Foucault Pendulum, 515
- Orkin, M. and R. Kakigi, What is the Worth of Free Casino Credit?, 3
- Östergård, P. *see Hämäläinen*
- P. Petersen *see Liggett*
- Peterson, B. E. and J. H. Jordan, Integer Hexahedra Equivalent to Perfect Boxes, 41
- Poole, D. G., The Stochastic Group, 798
- Rabinowitz, S. and S. Wagon, A Spigot Algorithm for the Digits of π , S. Rabinowitz and S. Wagon, 195
- Ranum, D. L., On Some Applications of Fibonacci Numbers, 640
- Redfield, R. H., Exponentiation in Rings, 36
- Robbins, D. P., Areas of Polygons Inscribed in a Circle, 523
- Robins, S. *see Diaz*
- Rosen, M. I., Niels Hendrik Abel and Equations of the Fifth Degree, 495
- Rudakov, A. N. *see Bashmakova*
- Sabinin, P. and M. G. Stone, Transforming n -gons by Folding the Plane, 620
- Sahoo, P. K., Three Open Problems in Functional Equations, 741
- Sandell, D. *see Blom*
- Sander, J. W., A Story of Binomial Coefficients and Primes, 802
- Scavo, T. R. and J. B. Thoo, On the Geometry of Halley's Method, 417
- Schay, G., How to Add Fast --- On Average, 725
- Schmeichel, E. *see Hausrath*
- Schneider, F. B. *see Gries*
- Scofield, P. D., Curves of Constant Precession, 531

- Shenitzer, A. *see* Mykytiuk
 Singer, D. A., Isometries of the Plane, 628
 Stillwell, J., Elliptic Curves, 831
 Stolarsky, K. B., Searching for Common Generalizations: The Case of Hyperbolic Functions, 609
 Stone, M. G. *see* Sabinin
 Taylor, A. D. *see* Brams
 Thoo, J. B. *see* Scavo
 Tokarsky, G. W., Polygonal Rooms Not Illuminated from Every Point, 867
 Toy, P. and S. Wagon, Calculus in the Operating Room, 101
 Van Der Poorten, A. J. *see* Myerson
 Van Wyk, C. J., Missing Real Numbers, 260
 Wagon, S. *see* Rabinowitz
 Wagon, S. *see* Toy
 Weintraub, S. H., Count-Wheels: A Mathematical Problem Arising in Horology, 310
 Wells, C., Communicating Mathematics: Useful Ideas from Computer Science, 397
 Weyl, H., Topology and Abstract Algebra as Two Roads of Mathematical Comprehension, Part I, 453
 Weyl, H., Topology and Abstract Algebra as Two Roads of Mathematical Comprehension, Part II, 646
 Zabell, S. L., Alan Turing and the Central Limit Theorem, 483
 Zulauf, A. *see* Flath

NOTES TITLE INDEX

- Adding Distinct Congruence Classes Modulo a Prime, N. Alon, M. B. Nathanson, and I. Ruzsa, 250
 Answers to Two Questions Concerning Quotients of Primes, P. Starni, 347
 Avoiding the Exchange Lemma, J. Ford, 350
 Calculating Normal Probabilities, R. J. Bagby, 46
 The Color Invariant for Knots and Links, P. Andersson, 442
 A Cone Eversion, S. Tabachnikov, 52
 Congruences Relating the Order of a Group to the Number of Conjugacy Classes, B. Poonen, 440
 Constrained Critical Points, P. Shutler, 49
 A Converse to Cauchy's Inequality, D. Zagier, 919
 The Derivative of the Exponential Map of Matrices, G. M. Tuynman, 818
 An Elementary Proof of the Simplicity of the Mathieu Groups M_{11} and M_{23} , R. J. Chapman, 544
 Entire Functions Which Vanish at Infinity, R. B. Burckel, 916
 Fibonacci-Like Sequences and Greatest Common Divisors, H. R. Morton, 731
 The Four-Vertex Theorem Revisited --- Two Variations on the Old Theme, S. Tabachnikov, 912
 Generating Symmetric Groups, I. M. Isaacs and T. Zieschang, 734
 An Inductive Proof of a Mixed Arithmetic-Geometric Mean Inequality, T. Matsuda, 634
 Injective Polynomial Maps are Automorphisms, W. Rudin, 540
 Intervals Contained in Arithmetic Combinations of Sets, S. Silverman, 351
 The Kantorovich Inequality, V. Pták, 820
 Matrix Expansion by Orthogonal Kronecker Products, J. C. Allen, 538
 More on Kummer's Test, H. Samelson, 817
 A Note on Entire Solutions of the Eiconal Equation, D. Khavinson, 159
 On the Arithmetic-Geometric Mean Inequality, L. G. Lucht, 739
 On the Generalized Inverse Form of the Equations of Constrained Motion, R. Kalaba and R. Xu, 821
 One More Construction Which is Impossible, V.A. Geyler, 632
 Permutations as Products of Transpositions, G. Mackiw, 438
 The Ranks of Tournament Matrices, T. S. Michael, 637
 A Relation Between Partitions and the Number of Divisors, W. Z. Bing, R. Fokkink, and W. Fokkink, 345
 A Short Path to the Shortest Path, P. D. Lax, 158
 A Simple Proof of the Hölder and the Minkowski Inequality, L. Maligranda, 256
 The Uniqueness Aspect of the Fundamental Theorem of Finite Abelian Groups, D. B. Surowski, 162
 Where Not to Find the Critical Points of a Polynomial --- Variation on a Putnam Theme, P. Andrews, 155

NOTES AUTHOR INDEX

- Allen, J. C., Matrix Expansion by Orthogonal Kronecker Products, 538
 Alon, N., M. B. Nathanson, and I. Ruzsa, Adding Distinct Congruence Classes Modulo a Prime, 250
 Andersson, P., The Color Invariant for Knots and Links, 442
 Andrews, P., Where Not to Find the Critical Points of a Polynomial --- Variation on a Putnam Theme, 155
 Bagby, R. J., Calculating Normal Probabilities, 46
 Bing, W. Z., R. Fokkink, and W. Fokkink, A Relation Between Partitions and the Number of Divisors, 345
 Burckel, R. B., Entire Functions Which Vanish at Infinity, 916
 Chapman, R. J., An Elementary Proof of the Simplicity of the Mathieu Groups M_{11} and M_{23} , 544
 Fokkink, R. *see* Bing
 Fokkink, W. *see* Bing
 Ford, J., Avoiding the Exchange Lemma, 350
 Geyler, V. A., One More Construction Which is Impossible, 632
 Isaacs, I. M. and T. Zieschang, Generating Symmetric Groups, 734
 Kalaba, R. and R. Xu, On the Generalized Inverse Form of the Equations of Constrained Motion, 821
 Khavinson, D., A Note on Entire Solutions of the Eiconal Equation, 159
 Lax, P. D., A Short Path to the Shortest Path, 158
 Lucht, L. G., On the Arithmetic-Geometric Mean Inequality, 739
 Mackiw, G., Permutations as Products of Transpositions, 438
 Maligranda, L., A Simple Proof of the Hölder and the Minkowski Inequality, 256
 Matsuda, T., An Inductive Proof of a Mixed Arithmetic-Geometric Mean Inequality, 634
 Michael, T. S., The Ranks of Tournament Matrices, 637
 Morton, H. R., Fibonacci-Like Sequences and Greatest Common Divisors, 731
 Nathanson, M. B. *see* Alon
 Poonen, B., Congruences Relating the Order of a Group to the Number of Conjugacy Classes, 440
 Pták, V., The Kantorovich Inequality, 820
 Rudin, W., Injective Polynomial Maps are Automorphisms, 540
 Ruzsa, I. *see* Alon
 Samelson, H., More On Kummer's Test, 817
 Shutler, P., Constrained Critical Points, 49
 Silverman, S., Intervals Contained in Arithmetic Combinations of Sets, 351
 Starni, P., Answers to Two Questions Concerning Quotients of Primes, 347
 Surowski, D. B., The Uniqueness Aspect of the Fundamental Theorem of Finite Abelian Groups, 162
 Tabachnikov, S., The Four-Vertex Theorem Revisited --- Two Variations on the Old Theme, 912
 Tabachnikov, S., A Cone Eversion, 52
 Tuynman, G. M., The Derivative of the Exponential Map of Matrices, 818
 Xu, R. *see* Kalaba
 Zagier, D., A Converse to Cauchy's Inequality, 919
 Zieschang, T. *see* Isaacs

REVIEWS BY TITLE

Names of authors are in ordinary type; those of reviewers in capitals.

- A Radical Approach to Real Analysis*, David Bressoud, SANDY GRABINER, 661
Algebra, I. M. Gelfand and Alexander Shen, RICHARD ASKEY, 78
Computability, Douglas S. Bridges, YIANNIS N. MOSCHOVAKIS, 752
Essays in Humanistic Mathematics, edited by Alvin White, ERIC LIVINGSTON, 846
Hilbert's Tenth Problem, Yuri V. Matiyasevich, MARTIN DAVIS, 366
Knot Theory, Charles Livingston, JOAN S. BIRMAN, 755
Modern Differential Geometry of Curves and Surfaces, Alfred Gray, BRUCE SOLOMON, 937
Politics, Logic, and Love: The Life of Jean van Heijenoort, Anita Burdman Feferman, JEFFREY NUNEMACHER, 178
Squares, A. R. Rajwade, DANIEL B. SHAPIRO, 281
The Words of Mathematics: An Etymological Dictionary of Mathematical Terms Used in English, Steven Schwartzman, HENRY J. RICARDO, 563

SOLUTIONS

Numbers in boldface refer to problems; those in lightface to pages.

10186	656	10248*	929	10269	74	10291	176
10201	929	10251	465	10270	558	10293	278
10206	657	10252	276	10272	560	10297	279
10208	361	10253	277	10273	748	10337	659
10213	172	10254	363	10274	658	10347	844
10222	173	10258	747	10275	76	10373	929
10230	72	10259	556	10276	750	6667*	929
10231	175	10260	364	10277	751	6668	464
10233	274	10262	467	10280	561	E 3473	171
10234	362	10264	365	10282	468	E3470*	929
10236	73	10265	842	10284	843		
10238	275	10267	748	10287	929		
10248	554	10268	557	10290	929		

*Revivals

PROBLEMS PROPOSED

Alkan, Emre 274, 745, 841	Cossi, Ernesto Bruno 274	Kaplansky, Irving <i>see</i> <i>Elkies</i>
Andersen, E. Sparre and	Darling, Donald A. 170	Kiechle, Hubert 359
Mogens Esrom Larsen 654	Davidson, Kenneth R. <i>see</i>	King, Jonathan L. 170
Anglesio, Jean 655	<i>Brown</i>	Kitchen, Edward 655
AT&T Bell Laboratories <i>see</i>	Diamond, Harold G. <i>see</i>	Klamkin, Murray S. 463,
<i>Rojas</i>	<i>Bishop</i>	929
Bang, Seung-Jin 463, 929	Duke, William 929	Klamkin, Murray S. <i>see</i>
Barnes, Allen 70	Elkies, Noam and Irving	<i>Alkan (841)</i>
Baumgartner, James E. and	Kaplansky 70	Kløve, Torleiv 553
Benjamin J. Tilly 655	Franco, Zachary 464	Knuth, Donald E. 655
Bebiano, N. 841	Galvin, Fred and John Isbell	Lagarias, Jeffrey C. 746
Beckwith, David 553	71	Larsen, Mogens Esrom <i>see</i>
Bielawski, Roger 274	Gauchman, Hillel and Lee A.	<i>Andersen</i>
Bishop, Richard L. and Harold	Rubel 554	Locke, Stephen C. 360
G. Diamond 274	Gessel, Ira 70	Neumann, B. H. 746
Bloom, David M. 169	Goffinet, Daniel 170	Poonen, Bjorn <i>see</i> <i>Lagarias</i>
Bradley, David 841	Grivaux, Jean-Pierre 929	Providência, J. da <i>see</i>
Brown, Daniel R. L., Kenneth	Gross, Alan J. and Hong	<i>Bebiano</i>
R. Davidson, and Jeffrey	Zhang 359	Rabinowitz, Stanley 841
Shallit 170	Hershkorn, Stephen J. 554	Reutenauer, Christophe <i>see</i>
Cater, F. S. 554	Higgins, Joseph E. 654	<i>Stanley</i>
Cavachi, Marius 273	Hutchinson, Joan P. 746	Rey, Joaquín Gómez 360
Chang, Fu-Chuen 360	Ionin, Yury J. 169	Rivin, Igor 554
Chao, Wu Wei 746	Isbell, John <i>see</i> <i>Galvin</i>	Rogers, D. G. <i>see</i> <i>Neumann</i>
Cohn, Henry 464	Just, Erwin 71	Rojas, J. Maurice and AT&T

- Bell Laboratories 170
 Rosset, Shmuel 840
 Rubel, Lee A. *see Gauchman*
 Schmidt, Frank and Louis W. Shapiro 464
 Schmidt, Frank 360, 840, 929
 Semmes, Stephen and Richard Stong 655
 Shallit, Jeffrey *see Brown*
 Shapiro, Daniel B. 464
 Shapiro, Louis W. *see Schmidt*
 Silverman, Joseph H. 841
 Snevily, Hunter S. 273
 Soules, George 71
 Spearman, Blair K. *see Williams*
 Stanley, Richard P. 929
 Stefanov, Simeon T. 746
 Stockmeyer, Paul K. 554
 Stong, Richard *see Semmes*
 Tamvakis, Harry 463, 745
 Tilly, Benjamin J. *see Baumgartner*
 Vanden Eynden, Charles 273
 Wardlaw, William P. 929
 Williams, Kenneth S. and Blair K. Spearman 360
 Wimp, Jet 71
 Zhang, Hong *see Gross*

PROBLEMS SOLVED

- Anchorage Math Solutions Group, 172, 277, 558
 Ash, J. Marshall 560
 Bailey, Duane W. 362
 Barbara, Roy 76
 Beckwith, David 175
 Binz, J. C. 365
 Bühler, Wolfgang J. 929
 Cantor, David G. 842
 Chapman, Robin J. 74, 172, 174, 465, 557, 561, 657, 747, 750
 Cohen, Graeme L. *see Iannucci*
 Cooper, Curtis 557
 Darling, Donald A. 842
 Drnovšek, Roman 275
 Dookovif, Dragomir D. 464
 Egerland, W. O. 277
 Fein, Burt 658
 Ford, Kevin 361
 Foregger, Thomas H. 362
 Gauchman, Hillel 844
 Griffin, Peter 929
 Hansen, C. E. *see Egerland*
 Hernández, Victor 555
 High, Robert 276
 Holzsager, Richard 175, 274, 555
 Iannucci, Douglas 72
 Izotov, Anatoly S. 467
 Kastanas, Ilias 929
 Klamkin, Murray S. 363
 Krop, Leonid *see Ash*
 LaBerge, Timothy J. 749
 Lindsey II, John H. 364
 Müller, Andreas 362
 National Security Agency Problems Group 559, 560, 748
 Nijenhuis, Albert 76
 Pedersen, Allan 561
 Pelling, M. J. 73
 Robinson, Raphael M. 278
 Robson, Robby *see Fein*
 Rosenholtz, Ira *see Gauchman*
 Ruderman, Harry D. 468
 Schilling, Kenneth 929
 Schmidt, F. 559
 Schmidt, Frank 176
 Stong, Richard 279
 University of Wyoming Problem Circle 656
 University of Wyoming Problem Circle *see Izotov*
 Vélez, Ricardo *see Hernández*
 Venkatachala, B. J. 468
 Vowe, Michael 659
 Wallen, Lawrence J. 171
 Wong, Yan Loi 556
 Woord, A. N. 't 278, 751

THANKS

The Monthly expresses its appreciation to the following people for their help in refereeing during the past year. We could not function without such people and their hard work.

Ralph Alexander, Richard Arratia, David Auckly, Sheldon Axler, Tom Banchoff, Paul T. Bateman, Bonnie Bennett, David Beyer, Patrick P. Billingsley, Andrew Bremner, David Bressoud, Joe Buhler, Robert B. Burckel, Bradley Carlin, Gulbank D. Chakerian, John B. Conway, John H. Conway, Robert M. Corless, David A. Cox, H. S. MacDonald Coxeter, Peter G. Doyle, Richard M. Dudley, William W. Dunham, John R. Durbin, Alan Durfee, Gerald A. Edgar, Allan Edmonds, Michael Filaseta, Dan Flath, Edward Frees, William Fulton, Richard J. Gardner, Betty Garrison, Leonard Gillman, Herman Gluck, Andrew Granville, David S. Griffeath, Branko Grünbaum, Denny Gulick, Leon Hall, Morris W. Hirsch, R. Daniel Hurwitz, Norman W. Johnson, David Lewis Johnson, Victor Katz, Louis H. Kauffman, Kevin P. Keating, Steven George Krantz, Joan Krone, Robert Barnard Kusner, Jeffrey C. Lagarias, Thomas M. Liggett, Douglas A. Lind, Charles Livingston, Michael Luby, Joseph Malkevitch, George E. Martin, Michal Misiurewicz, Hugh L. Montgomery, Alec Norton, Robert Osserman, Edward W. Packel, Jean Pedersen, Robin A. Pemantle, Ron Perline, Michael D. Perlman, George M. Phillips, Carl Pomerance, Stephen Portnoy, James Gary Propp, Leonard F. Richardson, Daniel G. Rider, John F. Rigby, Margaret Maher Robinson, Robby Robson, Kenneth H. Rosen, Richard Roth, Paula A. Russo, Doris W. Schattschneider, Daniel A. Schwalbe, Brigitte Servatius, Thomas Q. Sibley, Andrew F. Siegel, Richard P. Stanley, Peter Sternberg, Gilbert Strang, Lajos F. Takács, Jeremy T. Teitelbaum, Blake Temple, Ronald A. Thisted, Jeff David Vaaler, Richard S. Varga, J. B. Wilker, David E. Zitarelli.